Assessing selective processes acting on allele frequencies, genomic organisation, gene expression profiles in eukaryotic genomes

Wang, Wei

Award date:
2015

Awarding institution:
University of Bath

Assessing selective processes acting on allele frequencies, genomic organisation, gene expression profiles in eukaryotic genomes

Wei Wang

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

October 2014

COPYRIGHT

Table of contents

Acknowledgements

I would like to express my extreme gratitude and appreciation to my supervisor, Dr AraxiUrrutia, for her patience, guidance and support on both an academic and a personal level. This thesis would not have been possible without her advice and help.

I am most grateful to Dr Paula Kover for providing me with the experiment data and manuscript of her unpublished version of Arabidopsis flowering time.

I would like to thank Dr Elaine Wilkin, Dr Steve Dorus, Prof. Laurence Hurst, Dr Humberto Gutierrez, Steve Bush, Atahualpa Castillo-Morales, JimenaMonzon-Sandoval, Jaime Tovar-Corona and Qiong Wu for their advice, support and helpful discussion during my research studies. I would also like to thank Nina, Lu and members of Araxi's lab.

I would like to acknowledge the financial supports from my supervisor and from Department of Biology and Biochemistry of the university.

My special gratitude goes to Dr Guang-Zhong Wang, who introduced Bath to me, the city I fell in love at first sight.

Last, but certainly not least, I thank my parents for their material and spiritual support throughout my four years PhD study. Without their encouragement and love I never could have come this far.

Contributions

I acknowledge the following specific contributions:

1. The assessment of the significance of clustering of primate genes presented in Chapter 2 was jointly carried out with JimenaMonzon.

2. Clustering data used in Chapter 3 was obtained and kindly shared with me by Elaine Wilkins and Steve Dorus as part of a collaborative project.

3. Results regarding the significance of changes in allele frequencies for individual SNPs presented in Chapter 4 were kindly provided by Paula Kover as part of a collaboration.

All chapters presented were developed with advice from my supervisor. I also received valuable comments and suggestions on versions for one or more chapters from current and former PhD students in my group JimenaMonzon, Atahualpa Castillo, Stephen Bush and Jaime Tovar. Chapters 3 and 4 where prepared as part of collaborations with Steve Dorus and Paula Kover, respectively who participated in discussions regarding the design of each project.

Abstract

How genomes evolve through time and how they change in response to selective pressures remains a key area of research in genomics and evolutionary biology. Genome organization is now known to play a significant role in the regulation of expression patterns with significant clustering according to various parameters of gene expression having been reported in all major taxa.

In this thesis I present a comprehensive analysis of sex-biased gene expression in the primate genome and show that there is a significant degree of similarity in sex-biased gene expression among neighbouring genes. Whether this clustering of genes with similar expression profiles is functional or instead the result of transcriptional interference with adjacent genes displaying non-functional but significant similarity in patterns of gene expression has only recently started to be addressed. A recent study suggested that although *Drosophila melanogaster* exhibits significant similarities in gene expression among neighbouring genes, these clusters are not conserved across evolutionary time in the *Drosophila* lineage. Chapter three of this thesis presents a comprehensive analysis of the conservation of testis overexpressed gene clusters. I show that, as has been found for other *Drosophila* expression clusters, testis overexpression clusters are also not conserved throughout evolution. Finally, in chapter 4 I present an analysis of allele frequency changes in the *Arabidopsis thaliana* genome in experimentally selected lines for early flowering under two different growth conditions resembling winter and spring growth seasons. My results reveal widespread changes in allele frequencies in response to selective pressures with a significant degree of parallel changes for independent lines under selection in similar growth conditions. Importantly, from the point of view of

conservation crop efficiency efforts, no significant parallel changes were observed when examining the similarity in allele frequency changes across both growth conditions suggesting that adaptation for a particular trait might be only relevant in specific environmental conditions. Together these observations suggest that allele frequencies change on a global scale in response to selective pressures and that while the observed changes are mirrored across replicas in similar environments this is not the case for lines selected under different growth conditions.

Overall, the results I present in this thesis provide valuable insights into how gene order relates to gene expression profiles and its functional relevance as well as presenting evidence for patterns of allele changes in response to selective pressures.

Chapter 1 Introduction

To understand the genetic basis of how species adapt to changes in their living environment is one of the basic objectives in evolutionary biology. How genomes evolve through time and how they change in response to selective pressures are long standing questions in genomics and evolutionary biology. By analysing genome sequences in relation to functional variables such as gene expression patterns or functional annotations, numerous studies have built our understanding of the basic principles of how genes and genomes are organized. One of the major discoveries after the advent of whole genome sequencing was the clustering of genes along chromosomes in accordance with gene expression patterns (Hurst *et al*., 2004). Non-random gene order has been identified in a variety of different taxa (Price *et al*., 2006, Williams and Bowles, 2004, Lercher *et al*., 2002, Spellman and Rubin, 2002).

Genomes are in a constant evolution with new mutations arising constantly as a result from DNA repair mechanisms, DNA replication and transcription of gene expression. Most of the variation at the DNA level among individuals of the same species as well as fixed substitutions between species is considered to be neutral as proposed by Kimura(1983).

1.1 Next generation sequencing / RNA-seq

Thanks to the rapid development of genomic technologies, especially the advances in DNA sequencing techniques, it is, at this point in time, more practical to assess the genetic loci which may contribute to adaptive evolution. Using deep-sequencing technologies, RNA-sequencing (RNA-seq for short) provides us with a

much more extensive measurement of transcriptome than previous sequencing methods (Table 1-1)(Wang *et al*., 2009).

Table 1-1. Advantages of RNA-Seq compared with other transcriptomics methods. Taken from Wang *et al*. (2009)

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

In three self-contained studies presented in chapters 2-4 I explore patterns of genome organization and the conservation of gene synteny as well as analyse the patterns of allele change in response to selective pressures that together aim to contribute to our overall understanding of the genomic forces shaping how genomes evolve and respond to a changing environment.

1.2 Sex-biased gene expression characterization in primate species

In chapter 2, I assess patterns of sex-biased gene expression in primate genomes. Sexual dimorphism is the differences exhibited between males and females within a species. The phenotypic differences are often dramatic between males and females and can involve behaviour, morphology and physiology, including traits

such as the divergence of gametes and gonads, body size, coloration, etc(Mank, 2009).

To explain the exaggerated traits such as ornaments, coloration and singing which evolved in only one sex of a species, Darwin proposed the hypothesis of sexual selection (Kuijper *et al*., 2012). The competition between individuals (often males for most species) for the chance to mate and the opportunity to produce offspring results in striking sexual selection.

The majority of sexual dimorphisms are heritable suggesting a set sexual differentiation programme along development (Mank, 2009). An intriguing question raised by sexual dimorphism is what, at the molecular level, underlies the sometimes dramatic differences between females and males by using one almost identical genome? Several studies have shown that many sexual dimorphisms are manipulated by androgen or estrogen(Zauner *et al*., 2003, Mank *et al*., 2007, Ketterson *et al*., 2005). However, at the most basic level sexually dimorphic traits, even those mediated by hormonal responses, are the product of or are caused by preferential expression of a set of genes in one sex during embryonic, juvenile and adult development (Williams and Carroll, 2009).

The origination of sex-biased genes is hypothesized in at least three ways: single-locus sexual antagonism (intergenomic conflict), gene duplication, and non-coding sequence. The sexual antagonism hypothesis seems to be supported by the distribution of both female- and male-biased expression genes. Female-biased genes showed an excess of X-linkage, whereas male-biased genes are under-represented on the X chromosome(Connallon and Knowles, 2005). Duplication of a gene from an existing sex-biased gene represents a straightforward approach to introduce a new

sex-biased gene (Ellegren and Parsch, 2007). Evidence from both the worm (Cutter and Ward, 2005) and *Drosophila*(Gnad and Parsch, 2006) genomes demonstrates an increased number of male-biased genes through duplication. DNA sequences with no coding function could also provide a source for sex-biased expression gene generation. Levine *et al*. (2006) performed a whole-genome study of *Drosophila melanogaster* and identified five novel testis expressed genes which originated from ancestral non-coding sequence. Analysing the ESTs data from *Drosophila yakuba*, Begun *et al.*(2007) identified several *de novo* genes which were X-linked and exhibited testis-biased expression.

Sex determination mechanisms can be controlled by environmental cues as is the case in many fish and reptile species or be genetically determined by one or more genes located on either sex chromosomes or on an autosome (reviewed in Graves, 2006). Among species with genetically predetermined sex, many have developed specialized sex chromosomes. There are several common systems of sex chromosomes. In a ZW system, as observed in snakes, birds and butterflies, organisms exhibitheterogamety in females, which produces ZZ in males and ZW in females. In a XY system such as that observed in *Drosophila* and mammals, the males are heterogamous, which results in a XX karyotype in females and XY in males. Several studies have revealed that sex chromosome linked genes differ from autosomal genes in several ways. Previous studies have found evidence for the faster evolution of X-linked genes in mammals (Khaitovich *et al*., 2005) and Z-linked genes in birds (Mank *et al*., 2007). The fast-X effect is proposed to play a role in the evolution of genes harbouring on the X chromosome, in which the adaptive evolution rate was higher for X-linked genes than for autosomal genes (Baines *et al*., 2008).

A recent study showed a significant overrepresentation of female-biased genes residing on X chromosomes in both the *Drosophila* and mouse (Meisel *et al*., 2012). However, other species-specific studies do not reveal an extensive effect for sex chromosomes, compared to their proportion in genome (Ritchie, 2000, Wolfenbarger and Wilkinson, 2001). In a study of *Drosophila*, no evidence indicates an extra contribution of the X chromosome after a whole genome scan of *Drosophila*(Fitzpatrick, 2004).

These findings suggest that while there might be some over-representation of sex-biased genes in sex chromosomes this is not universal. On the most fundamental level, sex-determining or related genes are harboured on sex chromosomes, and trigger gonad differentiation. Sex determining genes located in sex chromosomes can then induce sex biased expression in downstream genes located in sex or autosomal chromosomes which ultimately results in the observed dimorphism. Although sex chromosomes contribute more to the evolution of sexual dimorphism compared to their relative physical size orgenic content, the effects of sex chromosomes are no more than of the components of autosomes (Mank, 2009). Many species which completely lack sex chromosomes, or have no sex-determining genes show noticeable sexual dimorphism (reviewed in Mank, 2009). This evidence further suggestsa major role of autosomal linked genes in encoding sexual dimorphism.

Thanks to the rapid development in genomics, the genetic mechanisms underlying sexual dimorphism have been revealed by several recent studies(Parisi *et al*., 2004, Yang *et al*., 2006, Mank *et al*., 2008, Reinius *et al*., 2008). Genomic evidence indicates that the divergence of female and male forms may be ascribed to thousands of genes, which show sex-biased expression, distributed across the

genome(Parisi *et al*., 2004, Yang *et al*., 2006, Mank *et al*., 2008, Reinius *et al*., 2008). A comparative genomic study in *Drosophila*(Zhang *et al*., 2004) shows that the evolution rate in sex-related traits, especially the traits related to male reproduction,is faster than that of non-sex-related traits. In contrast to female-biased genes or unbiased genes, male-biased genes show a significantly accelerated rate of evolution.

Sex is a major contributor to gene expression differences in a wide variety of animals. Sex-biased genes are likely to play an important role in sexual selection and speciation (Naurin *et al*., 2011) and changes in sex-biased gene expression are therefore likely to be a major contributor to adaptive phenotypic divergence between species (Ellegren and Parsch, 2007).

In chapter 2, I characterised sex biased gene expression in six primate species, and found that there is a tendency across most species for higher levels of male-biased expression and that higher levels of sex biased expression are associated with lower expression and higher tissue specificity and faster rates of protein evolution.

1.3 Genome organization and gene order evolution in eukaryotes

It is now widely recognized that in eukaryotic genomes, genes are distributed non-randomly across chromosomes according to their expression (Hurst *et al*., 2004, Lercher *et al*., 2002) and intensity (Caron *et al*., 2001). For example, in mammalian genomes, housekeeping genes, which are expressed in most tissues show a strong clustering in the human genome (Lercher *et al*., 2002). In *Drosophila* species,

thousands of testis specific genes were identified and observed to be clustered (Boutanaev *et al*., 2002).Numerous studies have found evidence that the loci that are most strongly differentiated between populations are sometimes clustered together, in what have been termed "genomic islands of divergence"(Turner *et al*., 2005, Via and West, 2008, Nadeau *et al*., 2012, Feder *et al*., 2012, Rogers *et al*., 2013).

Several evolutionary interpretations have been proposed in order to unfold the driving forces responsible for the clustering of functional neighbouring genes, including genomic rearrangement, chromosome inversion, (small-scale) transposition, and insertion and deletion (indel).

Larkin *et al*. (2009) performed an intriguing study by comparing genomes of amniotes and discovered that the breakpoints of rearrangements had a tendency to be clustered in the regions containing muscular contraction and inflammation related genes, while the conserved regions contained few rearrangement breakpoints with genes relating to development.

Therefore, rearrangements may have a tendency to be involved with local adaptation. Genomic rearrangement may compose an important part in the evolution of local adaptation and genomic divergence. If there is a sufficient population size, rate of rearrangement, selection pressure and migration rate, there will be rapid evolution of a highly clustered genomic structure (Yeaman, 2013).

Under heterogeneous circumstances, the architecture of the genome, as well as the neighbours of functional genes, may be shaped by natural selection (Yeaman 2013).

In chapter 3, I investigate whether testis over-expressed gene clusters are conserved through evolution of the *Drosophila* genus. The results I obtained demonstrate that gene order among testis overexpressed genes evolves rapidly with linkage acquired later in evolution and higher linkage breaks for ancestrally linked testis genes.

1.4 Selection on allele frequency in the *Arabidopsis* genome

For different local environments, adaptations can be identified by comparing allele frequencies between populations based on environmental variables. Several studies in *Arabidopsis thaliana* indicate that variants associated with flowering time are correlated with latitude(Verrelli and Eanes, 2001, Caicedo *et al*., 2004, Stinchcombe *et al*., 2004). Eckert *et al*. (2010) used a genome-wide data set of single nucleotide polymorphisms genotyped across 3059 functional genes to study patterns of population structure and identified loci associated with aridity across the natural range of loblolly pine (*Pinustaeda*L.). Using data across 61 worldwide populations, Hancock *et al*(2011) studied the relationship between allele frequencies and climate at the genome-wide scale. They detected the genetic loci, which contributed to the adaptations, and found a group of candidate SNPs associated with climate variables.

In Chapter 4, I investigated molecular changes that mediate short-term response to selection for earlier flowering time in *Arabidopsis thaliana* under two environmental conditions, and found that a surprisingly large number of unlinked SNPs showed changes in allele frequency significantly larger than expected under drift alone. Most SNPs identified to have responded to selection were different

between the two selection environments, and many of the SNPs showed significant changes in allele frequency in a single selection line.

Together the results presented in this thesis represent an exploration of the changes in genome organization and allele frequency distributions in association with different aspects which can impact and constrain the evolutionary paths at a large scale.

1.5 References

BAINES, J. F., SAWYER, S. A., HARTL, D. L. & PARSCH, J. 2008. Effects of X-linkage and sex-biased gene expression on the rate of adaptive protein evolution in Drosophila. *Mol Biol Evol,* 25**,** 1639-50.

BEGUN, D. J., LINDFORS, H. A., KERN, A. D. & JONES, C. D. 2007. Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. *Genetics,* 176**,** 1131-7.

BOUTANAEV, A. M., KALMYKOVA, A. I., SHEVELYOV, Y. Y. & NURMINSKY, D. I. 2002. Large clusters of co-expressed genes in the Drosophila genome. *Nature,* 420**,** 666-9.

CAICEDO, A. L., STINCHCOMBE, J. R., OLSEN, K. M., SCHMITT, J. & PURUGGANAN, M. D. 2004. Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. *Proc Natl Acad Sci U S A,* 101**,** 15670-5.

CARON, H., VAN SCHAIK, B., VAN DER MEE, M., BAAS, F., RIGGINS, G., VAN SLUIS, P., HERMUS, M. C., VAN ASPEREN, R., BOON, K., VOUTE, P. A., HEISTERKAMP, S., VAN KAMPEN, A. & VERSTEEG, R. 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science,* 291**,** 1289-92.

CONNALLON, T. & KNOWLES, L. L. 2005. Intergenomic conflict revealed by patterns of sex-biased gene expression. *Trends in Genetics,* 21**,** 495-499.

CUTTER, A. D. & WARD, S. 2005. Sexual and temporal dynamics of molecular evolution in C. elegans development. *Mol Biol Evol,* 22**,** 178-88.

ECKERT, A. J., VAN HEERWAARDEN, J., WEGRZYN, J. L., NELSON, C. D., ROSS-IBARRA, J., GONZALEZ-MARTINEZ, S. C. & NEALE, D. B. 2010. Patterns of population structure and environmental associations to aridity across the range of loblolly pine (Pinus taeda L., Pinaceae). *Genetics,* 185**,** 969-82.

ELLEGREN, H. & PARSCH, J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet,* 8**,** 689-698.

FEDER, J. L., EGAN, S. P. & NOSIL, P. 2012. The genomics of speciation-with-gene-flow. *Trends in Genetics,* 28**,** 342-350.

FITZPATRICK, M. J. 2004. Pleiotropy and the Genomic Location of Sexually Selected Genes. *The American Naturalist,* 163**,** 800-808.

GNAD, F. & PARSCH, J. 2006. Sebida: a database for the functional and evolutionary analysis of genes with sex-biased expression. *Bioinformatics,* 22**,** 2577-9.

GRAVES, J. A. 2006. Sex chromosome specialization and degeneration in mammals. *Cell,* 124**,** 901-14.

HANCOCK, A. M., WITONSKY, D. B., ALKORTA-ARANBURU, G., BEALL, C. M., GEBREMEDHIN, A., SUKERNIK, R., UTERMANN, G., PRITCHARD, J. K., COOP, G. & DI RIENZO, A. 2011. Adaptations to climate-mediated selective pressures in humans. *PLoS Genet,* 7**,** e1001375.

HURST, L. D., PAL, C. & LERCHER, M. J. 2004. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet,* 5**,** 299-310.

KETTERSON, E. D. K., JR, V. N. & SANDELL, M. 2005. Testosterone in Females: Mediator of Adaptive Traits, Constraint on Sexual Dimorphism, or Both? *The American Naturalist,* 166**,** S85-S98.

KHAITOVICH, P., HELLMANN, I., ENARD, W., NOWICK, K., LEINWEBER, M., FRANZ, H., WEISS, G., LACHMANN, M. & PAABO, S. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science,* 309**,** 1850-4.

KIMURA, M. 1983. *The neutral theory of molecular evolution,* Cambridge Cambridgeshire ; New York, Cambridge University Press.

KUIJPER, B., PEN, I. & WEISSING, F. J. 2012. A Guide to Sexual Selection Theory. *Annual Review of Ecology, Evolution, and Systematics, Vol 43,* 43**,** 287-+.

LARKIN, D. M., PAPE, G., DONTHU, R., AUVIL, L., WELGE, M. & LEWIN, H. A. 2009. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Res,* 19**,** 770-7.

LERCHER, M. J., URRUTIA, A. O. & HURST, L. D. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet,* 31**,** 180-3.

LEVINE, M. T., JONES, C. D., KERN, A. D., LINDFORS, H. A. & BEGUN, D. J. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster

are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A,* 103**,** 9935-9.

MANK, J. E. 2009. Sex chromosomes and the evolution of sexual dimorphism: lessons from the genome. *Am Nat,* 173**,** 141-50.

MANK, J. E., AXELSSON, E. & ELLEGREN, H. 2007. Fast-X on the Z: rapid evolution of sex-linked genes in birds. *Genome Res,* 17**,** 618-24.

MANK, J. E., HULTIN-ROSENBERG, L., WEBSTER, M. T. & ELLEGREN, H. 2008. The unique genomic properties of sex-biased genes: insights from avian microarray data. *BMC Genomics,* 9**,** 148.

MEISEL, R. P., MALONE, J. H. & CLARK, A. G. 2012. Disentangling the relationship between sex-biased gene expression and X-linkage. *Genome Res,* 22**,** 1255-65.

NADEAU, N. J., WHIBLEY, A., JONES, R. T., DAVEY, J. W., DASMAHAPATRA, K. K., BAXTER, S. W., QUAIL, M. A., JORON, M., FFRENCH-CONSTANT, R. H., BLAXTER, M. L., MALLET, J. & JIGGINS, C. D. 2012. Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 367**,** 343-353.

NAURIN, S., HANSSON, B., HASSELQUIST, D., KIM, Y. H. & BENSCH, S. 2011. The sex-biased brain: sexual dimorphism in gene expression in two species of songbirds. *BMC Genomics,* 12**,** 37.

PARISI, M., NUTTALL, R., EDWARDS, P., MINOR, J., NAIMAN, D., LU, J., DOCTOLERO, M., VAINER, M., CHAN, C., MALLEY, J., EASTMAN, S. & OLIVER, B. 2004. A survey of ovary-, testis-, and soma-biased gene expression in Drosophila melanogaster adults. *Genome Biol,* 5**,** R40.

PRICE, M. N., ARKIN, A. P. & ALM, E. J. 2006. The life-cycle of operons. *PLoS Genet,* 2**,** e96.

REINIUS, B., SAETRE, P., LEONARD, J. A., BLEKHMAN, R., MERINO-MARTINEZ, R., GILAD, Y. & JAZIN, E. 2008. An evolutionarily conserved sexual signature in the primate brain. *PLoS Genet,* 4**,** e1000100.

RITCHIE, M. G. 2000. The inheritance of female preference functions in a mate recognition system. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 267**,** 327-332.

ROGERS, S. M., MEE, J. A. & BOWLES, E. 2013. The consequences of genomic architecture on ecological speciation in postglacial fishes. *Current Zoology,* 59**,** 53-71.

SPELLMAN, P. T. & RUBIN, G. M. 2002. Evidence for large domains of similarly expressed genes in the Drosophila genome. *J Biol,* 1**,** 5.

STINCHCOMBE, J. R., WEINIG, C., UNGERER, M., OLSEN, K. M., MAYS, C., HALLDORSDOTTIR, S. S., PURUGGANAN, M. D. & SCHMITT, J. 2004. A latitudinal cline in flowering time in Arabidopsis thaliana modulated by the flowering time gene FRIGIDA. *Proc Natl Acad Sci U S A,* 101**,** 4712-7.

TURNER, T. L., HAHN, M. W. & NUZHDIN, S. V. 2005. Genomic Islands of Speciation in <italic>Anopheles gambiae</italic>. *PLoS Biol,* 3**,** e285.

VERRELLI, B. C. & EANES, W. F. 2001. Clinal Variation for Amino Acid Polymorphisms at the Pgm Locus in Drosophila melanogaster. *Genetics,* 157**,** 1649-1663.

VIA, S. & WEST, J. 2008. The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology,* 17**,** 4334-4345.

WANG, Z., GERSTEIN, M. & SNYDER, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet,* 10**,** 57-63.

WILLIAMS, E. J. & BOWLES, D. J. 2004. Coexpression of neighboring genes in the genome of Arabidopsis thaliana. *Genome Res,* 14**,** 1060-7.

WILLIAMS, T. M. & CARROLL, S. B. 2009. Genetic and molecular insights into the development and evolution of sexual dimorphism. *Nat Rev Genet,* 10**,** 797-804.

WOLFENBARGER, L. L. & WILKINSON, G. S. 2001. SEX-LINKED EXPRESSION OF A SEXUALLY SELECTED TRAIT IN THE STALK-EYED FLY, CYRTODIOPSIS DALMANNI. *Evolution,* 55**,** 103-110.

YANG, X., SCHADT, E. E., WANG, S., WANG, H., ARNOLD, A. P., INGRAM-DRAKE, L., DRAKE, T. A. & LUSIS, A. J. 2006. Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res,* 16**,** 995-1004.

YEAMAN, S. 2013. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc Natl Acad Sci U S A,* 110**,** E1743-51.

ZAUNER, H., BEGEMANN, G., MARÍ-BEFFA, M. & MEYER, A. 2003. Differential regulation of msx genes in the development of the gonopodium,

an intromittent organ, and of the "sword," a sexually selected trait of swordtail fishes (Xiphophorus). *Evolution & Development,* 5**,** 466-477.

ZHANG, Z., HAMBUCH, T. M. & PARSCH, J. 2004. Molecular evolution of sex-biased genes in Drosophila. *Mol Biol Evol,* 21**,** 2130-2139.

Chapter 2 Characterisation of sex-biased genes in human and other primate genomes

2.1 Abstract

Sex-biased gene expression has been documented in many different taxa. Differences in gene expression patterns between males and females, particularly in the central nervous system, may underlie behavioural differences between the sexes. Here we characterise sex-biased gene expression in five tissues of six primate species, including humans. We find that, consistent with previous multispecies studies in *Drosophila*, there is a tendency across the primates for higher levels of male-biased expression, and that higher levels of sex-biased expression are associated with lower expression, higher tissue specificity and faster rates of protein evolution. Interestingly, we find that male-biased genes tend to cluster in specific regions within chromosomes and that, importantly, this pattern is not explained by the known clustering of testis-overexpressed genes along chromosomes. We further examined whether patterns of sex-biased gene expression were conserved across species, but we found little correlation in the degree of sex-biased gene expression between species in any tissue analysed suggesting that, as has been reported in previous studies, sex-biased gene expression is a rapidly evolving trait. We also found a significant and previously unreported clustering of sex-biased genes along chromosomes which suggests a strong influence of genomic neighbourhoods and chromatin structure in defining levels of sex-biased gene expression. Together these

results constitute a robust characterisation of sex-biased gene expression in the primate lineage.

2.2 Introduction

Many animal species show sexual dimorphism, with associated sex-biased differences in gene expression (Ellegren and Parsch, 2007). Although males and females share the vast majority of their genomes, sex-biased gene expression has been characterised for a diverse range of metazoans, including mosquitoes (Marinotti et al., 2006), mice (Yang et al., 2006), chickens (Mank et al., 2008, Ellegren et al., 2007), songbirds (Naurin et al., 2011), worms (Reinke et al., 2004), zebra fish (Small et al., 2009) and fruit flies (Parisi et al., 2003, Connallon and Clark, 2011, Assis et al., 2012, Jiang and Machado, 2009, Ranz et al., 2003, Perry et al., 2014, Zhang et al., 2004). From such studies, several characteristics of sex-biased genes (i.e. those with a female:male ratio of expression level deviating from zero) have been identified as common to multiple taxa, such as male-biased genes having comparatively elevated sequence divergence (Meiklejohn et al., 2003, Ranz et al., 2003, Zhang et al., 2004, Mank et al., 2007a, Mank et al., 2007b), reduced codon usage bias and a greater rate of gene turnover than either female-biased or unbiased genes (reviewed in Ellegren and Parsch, 2007). Genes with sexually dimorphic expression profiles are often expressed in specific tissues in both *Drosophila*(Meisel, et al. 2012) and mice (Yang, et al. 2006).

Studies of sex-biased gene expression in *Drosophila melanogaster* (Parisi, et al. 2004) and mice (Yang, et al. 2006) have shown that thousands of genes have significant levels of sex-biased gene expression. Selective explanations for

widespread sex-biased gene expression have emphasised the roles of sexual selection and sexual antagonism (Connallon and Knowles 2005; Proschel, et al. 2006; Cox and Calsbeek 2009) in the establishment and maintenance of a gene's sex-bias. Nevertheless, the set of sex-biased genes has been found to be highly variable among species (Metta, et al. 2006; Zhang, et al. 2007; Reinius, et al. 2008), even when considering relatively closely related species (Zhang, et al. 2007). Large variations in the set of genes exhibiting sex bias could be suggestive of highly species-specific selective pressures driving sex-biased gene expression, but it could equally reflect the non-specific co-regulation of large numbers of genes as a consequence of sex-specific variations in chromosomal packing. In particular, it is unclear as to whether male- or female-biased genes typically, or primarily, execute sex-specific functions, such as those involved in reproduction, or instead have similar functional importance for both sexes (Connallon and Clark 2011). Alternatively, widespread sex-biased gene expression may be explained by chromatin structure. In the human genome, genes have been shown to cluster along the genome according to their levels of gene expression (Caron, et al. 2001) and breadth of expression (Lercher, et al. 2002; Versteeg, et al. 2003). This non-random distribution of genes according to expression patterns might result from different processes (Hurst, et al. 2004). In mammals, chromatin structure variations along chromosomes might explain the higher than expected similarity in expression patterns among genes located within the same chromatin domain (Lercher et al. 2003). On a shorter range, the opening of chromatin to allow transcription of an individual gene can lead to the non-specific transcription of adjacent genes (Spellman and Rubin 2002). On the other hand, in mammals, adjacent genes may have similar functions and be co-transcribed accordingly – for instance, spermatogenesis genes accumulate in on the X

chromosome (Wang, et al. 2001; Lercher, et al. 2003; Mueller, et al. 2008). This excess of spermatogonial genes in on the X chromosome is limited to genes with pre-meiotic expression, whereas post-meiotically expressed genes are under-represented on the X chromosome because of male meiotic sex-chromosome inactivation (Khil, et al. 2004). Clustering of testis-overexpressed genes along chromosomes has also been reported in *Drosophila melanogaster* (Boutanaev, et al. 2002).

Sex-biased genes have been shown to be non-randomly distributed in the genome. In *Drosophila*, male-biased genes are underrepresented on the X chromosome (Parisi, et al. 2003; Ranz, et al. 2003; Sturgill, et al. 2007; Mikhaylova and Nurminsky 2011) with an excess of X-linked genes with female- and ovary-biased expression. Nevertheless, relatively new genes with male-biased expression do tend to be initially enriched on the X chromosome (Zhang, et al. 2010), but as gene age increases the proportion of male-biased X-linked genes reduces. In *Drosophila serrata*, the genomic location of sex-biased genes is already known to be non-random with an over-representation of female-biased genes on the X chromosome and a deficit of male X-linked genes (Allen, et al. 2013). By contrast, in a study of chicken sexual dimorphism, Mank et al. (2010) observed that most sex-biased genes reside on the autosomes.

Whether sex-biased genes in the primates are non-randomly distributed within, rather than among, chromosomes remains to be explored. Here we characterise sex-biased gene expression in five tissues for six primate species. We show that sex-biased gene expression tends to be tissue-specific, associated with low gene expression patterns and high rates of protein evolution. Interestingly, we

observed a significant similarity of sex-biased gene expression among adjacent genes. Moreover, this pattern is not explained by the known clustering of testes' over-expressed genes (Boutanaev, et al. 2002). Together these results constitute a robust characterisation of sex-biased gene expression in the primate lineage.

2.3 Materials and methods

2.3.1 Sex-biased gene expression level.

RNA-seq data for each of six tissues (brain, cerebellum, heart, liver, kidney, testis) in six adult primates (*Homo sapiens*, *Gorilla gorilla*, the rhesus macaque [*Macacamulatta*], the Borneanorangutan [*Pongopygmaeus*], the bonobo [*Pan paniscus*] and the common chimpanzee [*Pan troglodytes*]), were obtained from Brawand et al. (2011). With the exception of the testes, data were available for both males and females. Per-gene expression data were normalised against the total expression per tissue. Where possible, a ratio of female:male gene expression (F:M expression ratio) per gene pertissue and per species, was calculated (tissues available are summarized in Table S2-8). We excluded the orangutan cerebellum and the human liver as no data was available for these tissues. F:M expression ratios were $\log_2$-transformed to equate the scales for male and female biased expression analysis, as per Zhang et al. (2007).

2.3.2 Chromosomal distribution of sex-biased genes.

Gene positions for each species were obtained from EnsemblBioMart v73 (Kinsella et al., 2011). For each species except bonobo (genome assembly information not available at this stage), we calculated the average

female:maleexpression ratio per chromosome using sliding windows of variable lengths. Each window encompassed 50 genes, with length equal to the midpoint of the last gene minus the midpoint of the first gene. Windows moved along each chromosome gene-by-gene. We averaged the female:male expression ratios of all genes within the window.

2.3.3 Characteristics of sex-biased genes.

The tissue specificity index, *tau*, was calculated for each gene per species. *Tau* is a scalar measure of how broadly expressed a gene is, bounded between 0 (for housekeeping genes) and 1 (for genes expressed in one tissue only), defined as:

$$tau = \frac{\sum_{i=1}^{N}(1-x_i)}{N-1},$$

where*N* is the number of tissues and x_i is the expression profile component normalized by the maximal component value (Yanai et al., 2005). The degree and direction of selection acting on each gene was estimated using dN/dS, a standard measure of the rate of non-synonymous to synonymous changes per gene. For each gene in the six primate species, pairwise local alignments between the longest transcript of each gene and its corresponding mouse ortholog (sequences obtained from EnsemblBioMart (Kinsella, et al. 2011); only one-to-one orthologous were considered) were calculated using the Smith–Waterman algorithm (fasta36.3.5d with parameters –a –A) (Pearson 2000). Transcripts with an incomplete coding DNA sequences (CDS) were excluded from analysis. dN/dS was estimated using the Yang and Nielson model, as implemented in the yn00 package of PAML (Yang 2007).

2.3.4 Clustering of similarly sex-biased expression genes.

In order to test whether genes with the highest sex-bias tend to cluster within the chromosomes, we ranked all genes according to their F:M ratio. We focused on the top 10% of genes with the highest male-biased gene expression and those corresponding to the 10% with the highest female-biased gene expression. We defined clusters as a group of *n* adjacent genes, and counted the frequency and size of those containing genes within the top 10% of both the male and female distributions. We then compared the observed frequency to an expected frequency in equally sized clusters based on Monte Carlo simulations with 10,000 random samples drawn from the population of genesfor which expression data were available. Numeric p-values were calculated based on the 10,000 randomisations and then adjusted for multiple testing with the Bonferroni correction.

2.3.5 Similarity of sex biased expression in neighbouring genes.

We wanted to determine whether a gene with highly sex-biased expression tends to have neighbouring genes that are similarly sex-biased. We considered the midpoint location of each Ensembl gene ID simply as the average of gene start and gene end. We selected all possible gene pairs that were separated by a maximum of 10 genes based on the midpoint of each gene. We counted the gene pairs at different distances where both genes were highly sex-biased (either within the top 10% maleor female-biased) and compared this frequency to an expected number of neighbouring gene pairs drawn from 10,000 randomisations of the gene location.

2.3.6 General data analysis

All statistical analyses were performed in R.

2.4 RESULTS

2.4.1 Characterisation of sex-biased expression in six primate species

In order to examine sex-biased gene expression in five tissues of six primate species (human, bonobo, chimpanzee, gorilla, orangutan and the macaque), we first calculated sex-biased gene expression for each gene by obtaining the female to male expression ratio for six primate species using RNA-seq expression data derived from Brawand et al. (2011) (see Materials and Methods). We found strong variations in sex-biased expression distributions between tissues and between species (ANOVA test $F=462143$, $p < 0.001$, Table 2-2 and SupplementaryFigure 2-1, Supplementary Table S2-9 – S2-7).

Generally, we observed greater male-biased gene expression relative to female-biased gene expression, but this was not observed across all species: while in human, bonobo and chimp, the number of genes with a negative female to male expression ratio was significantly greater than genes with a positive ratio, in macaque the reverse was observed (Figure 2-1**Ошибка! Источник ссылки не найден.**; Table 2-2). Moreover, we also observed that, in general, genes with higher expression in males than in females tend to have higher deviations from female expression profiles. In other words, not only there is a tendency towards higher numbers of male-biased genes but as a group they tend to show higher degrees of sex-biased expression than those genes which are female-biased (Figure 2-1B).

Both of these findings are consistent with previous observations in the *Drosophila* lineage (Zhang et al., 2007) showing that most, but not all, species of the

seven examined displayed higher male-biased gene expression and that the magnitude of those biases tends to be higher for male-biased genes.

We then estimated tau, an index which measures tissue specificity of expression patterns (Chan et al., 2012) for each gene (Materials and Methods). We found that higher sex bias in the patterns of gene expression is associated with higher tissue specificity(Figure 2-2; Table 2-3).We further found that higher degrees of sex-biased expression are associated with lower expression levels (Figure 2-3,Table 2-4). This pattern was found to be consistent when examining individual tissues per species (Supplementary Figures S2-2 – S2-6, Supplementary Tables S2-8 – S2-9)

It has been reported that sex-biased genes, particularly those with a male expression bias, display high rates of protein evolution(Zhang et al., 2004, Ranz et al., 2003) which has been attributed to positive selection (Swanson and Vacquier, 2002, Zhang and Parsch, 2005). Male-biased genes, in particular, often show higher divergence rates between species (Ellegren and Parsch, 2007). In *Drosophila*, genes with male-biased expression, particularly those expressed in reproductive tissues, show consistently high rates of adaptive protein evolution (Zhang et al., 2004, Zhang et al., 2007, Proschel et al., 2006).

To assess whether this might be found between primate species, we calculated dN/dS ratios for each species against the corresponding mouse ortholog to ensure that all dN/dS values for genes in each primate species are calculated against a similarly divergent species and are thus comparable across species. We observed no significant associations between dN/dS and female:male expression ratioacross six primate species (Figure 2-4, Table 2-5, Supplementary figures S2-7 – S2-11, Supplementary table S2-2 –S2-10).

2.4.2 Clustering of sex-biased genes

Various studies have shown a non-random distribution of genes with sex-biased expression when comparing autosomes and sex chromosomes (Rice, 1984, Zhang et al., 2004, Ellegren and Parsch, 2007, Yang et al., 2006, Naurin et al., 2011). We analysed differences in a gene's female:male expression ratio in relation to its chromosomal location. Due to the lack of a complete genome assembly for the bonobo, it was excluded from this analysis.

We then explored the distribution of genes within chromosomes according to their degree of sex-biased expression. Allocating all the genes onto the chromosomes according to their genomic location, we assayed the distribution of male- and female-biased genes. For most of the primates, with the exception of gorilla, we observed significant deviations in the distribution of sex-biased expression among chromosomes (Figure 2-5, Table 2-6), which suggests the sex-biased expression is not randomly distributed across chromosomes. In addition, we observed a higher tendency for male-biased genes to cluster than for female-biased genes when using a sliding window scan analysis along each chromosome (Figure 2-6). The valleys and peaks in Figure 2-6 indicate that there may be clustering of male-biased or female-biased gene expression in these genomic regions.

In order to test whether the most biased genes tend to cluster along chromosomes, we selected the 10% of genes with the highest levels of male or female-biased gene expression. First, we obtained all contiguous gene pairs separated by different distances (measured as the number of genes between them) in the same chromosome. Then, we asked whether those gene pairs were similarly biased, that is, if they both belong to the top 10 % of genes with the highest female-

or male-biased expression. Finally, we tested for the enrichment of gene pairs with similarly biased expression compared to the rest of possible gene pairs at the same genomic distance. As shown in Figure 2-7 we found that sex-biased genes tend to be next to other similarly biased genes. As the distance between genes pairs increases the effect disappears– for gene pairs separated by more than three genes we no longer identify any enrichment. For macaque, chimp and orangutan the effect is stronger for the female-biased gene pairs (Figure 2-7A), while gorilla and human have more adjacent male-biased gene pairs than expected at the corresponding distance (Figure 2-7B).

Next, we asked whether these highest female- or male-biased genes tend to accumulate in a cluster. We calculated the number of different sized clusters of highly biased genes, and compared it to the expected number of equally sized clusters obtained through 10,000 randomizations. What we found is a significant enrichment of sex-biased gene clusters of at least three genes across fiveprimate species, except the male-biased genes in macaque and the female-biased genes in human (Figure 2-8).

2.4.3 Clustering of sex-biased genes is not explained by clustering of testis- overexpressed genes

One possible explanation of the within-chromosome clustering of male-biased genes could be the systemic over-expression, and non-random distribution, of genes with sex-specific functionality, such as those involved in the testes. To test this, we first calculated the $\log_2$ ratio of testis expression over the average expression in male across tissues to obtain an index of testis-overexpression per gene. When mapping gene positions along chromosomes according to their degree of sex-biased

expression and testis-overexpression we observed that the areas of male-biased gene expression do not match the clusters of testis over-expressed genes (Figure 2-6).A regression analyses showed that there is little relationship between the degree of sex-biased expression and testis over-expression in all species (Table 2-7). This suggests that the observed clustering of genes according of sex-biased gene expression is not explained by previously reported clustering of testis over-expression.

2.5 Discussion

A previous study in *Drosophila*(Zhang et al., 2007) observed greater expression of male-biased genes. Consistent with this finding, we found that in most species analysed, there is a higher number of genes with male-biased expression patterns than the number of female-biased genes. Furthermore, the average bias was found to be higher among male biased genes than the bias in female-biased genes in all primates except macaque.

Our analysis also shows a high degree of tissue specificity for sex-biased genes. These results are consistent with a previous study in *Drosophila*(Meisel et al., 2012) where sex-biased genes were found to be narrowly expressed in a limited number of tissues.

A recent study (Assis et al., 2012) has shown that in both *D. melanogaster* and *D. pseudoobscura*, female-biased genes tend to have smaller *tau* values than male-biased genes. However, in primate species, we did not find a consistent pattern of female-biased genes showing lower tissue specificity than male-biased genes. On the contrary, we found that as the strength of the sex bias increases, irrespective of the sex, so does the value of *tau*.

In a previous study, male-biased genes were found to display a strong and consistent signal of positive selection, while female-biased genes showed more variation in the type of selection they experience. Furthermore, genes expressed equally in the two sexes showed no evidence for adaptive evolution between *D. melanogaster* and *D. simulans*(Proschel et al., 2006). However, our observation in primates did not show any significant relationship between sex-biased expression and dN/dS.

In general, male-biased clusters were observed on both the X chromosome and autosomes in each primate species. Meanwhile, we also observed a gene neighbourhood effect in which two male-biased genes tend to be together in a greater frequency than expected by chance. This phenomenon might be partly explained by the clustering of testis genes along each chromosome. However, acorrelation analysis between sex-biased and testis-overexpressed gene clustersdoes not support this hypothesis, suggesting that other factors, such as small-scale rearrangement caused by positive selection, may affect the distribution of sex-biased genes along chromosomes.

The clustering of genes within chromosomes according to their expression profile has been observed in all major taxa (Caron et al., 2001, Boutanaev, et al. 2002), with the co-regulation of these clusters partly explained by broad transcriptional regulation acting on chromatin domains (Hurst et al., 2004, Kalmykova et al., 2005, Purmann et al., 2007, Branco et al., 2013). We investigatedthe within-chromosome distribution of sex-biased genes – those showing higher expression in one sex compared to the other – and found that there is significant clustering of both male and female-biased genes in several primates.

Importantly, this clustering is not explained by the known clustering of testis over-expressed genes. Taken together our results show extensive clustering of genes according to their degree of sex-biased expression which may resultfrom processes like sex-specific chromatin configurations or non-specific gene expression around a small set of genes with strong sex-biased expression.

2.6 References

ALLEN, S. L., BONDURIANSKY, R. & CHENOWETH, S. F. 2013. The genomic distribution of sex-biased genes in drosophila serrata: X chromosome demasculinization, feminization, and hyperexpression in both sexes. *Genome Biol Evol,* 5**,** 1986-94.

ASSIS, R., ZHOU, Q. & BACHTROG, D. 2012. Sex-biased transcriptome evolution in Drosophila. *Genome Biol Evol,* 4**,** 1189-200.

BOUTANAEV, A. M., KALMYKOVA, A. I., SHEVELYOV, Y. Y. & NURMINSKY, D. I. 2002. Large clusters of co-expressed genes in the Drosophila genome. *Nature,* 420**,** 666-9.

BRANCO, A. T., HARTL, D. L. & LEMOS, B. 2013. Chromatin-Associated Proteins HP1 and Mod(mdg4) Modify Y-Linked Regulatory Variation in the Drosophila Testis. Genetics, 194, 609-618.

BRAWAND, D., SOUMILLON, M., NECSULEA, A., JULIEN, P., CSARDI, G., HARRIGAN, P., WEIER, M., LIECHTI, A., AXIMU-PETRI, A., KIRCHER, M., ALBERT, F. W., ZELLER, U., KHAITOVICH, P., GRUTZNER, F., BERGMANN, S., NIELSEN, R., PAABO, S. & KAESSMANN, H. 2011. The evolution of gene expression levels in mammalian organs. *Nature,* 478**,** 343-8.

CARON, H., VAN SCHAIK, B., VAN DER MEE, M., BAAS, F., RIGGINS, G., VAN SLUIS, P., HERMUS, M. C., VAN ASPEREN, R., BOON, K., VOUTE, P. A., HEISTERKAMP, S., VAN KAMPEN, A. & VERSTEEG, R. 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science,* 291**,** 1289-92.

CHAN, YINGGUANG F., JONES, FELICITY C., MCCONNELL, E., BRYK, J., BÜNGER, L. & TAUTZ, D. 2012. Parallel Selection Mapping Using Artificially Selected Mice Reveals Body Weight Control Loci. *Current Biology,* 22**,** 794-800.

CONNALLON, T. & CLARK, A. G. 2011. Association between sex-biased gene expression and mutations with sex-specific phenotypic consequences in Drosophila. *Genome Biol Evol,* 3**,** 151-5.

CONNALLON, T. & KNOWLES, L. L. 2005. Intergenomic conflict revealed by patterns of sex-biased gene expression. *Trends Genet,* 21**,** 495-9.

COX, R. M. & CALSBEEK, R. 2009. Sexually antagonistic selection, sexual dimorphism, and the resolution of intralocus sexual conflict. *Am Nat,* 173**,** 176-87.

ELLEGREN, H., HULTIN-ROSENBERG, L., BRUNSTROM, B., DENCKER, L., KULTIMA, K. & SCHOLZ, B. 2007. Faced with inequality: chicken do not have a general dosage compensation of sex-linked genes. *BMC Biol,* 5**,** 40.

ELLEGREN, H. & PARSCH, J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet,* 8**,** 689-98.

HURST, L. D., PAL, C. & LERCHER, M. J. 2004. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet,* 5**,** 299-310.

JIANG, Z. F. & MACHADO, C. A. 2009. Evolution of sex-dependent gene expression in three recently diverged species of Drosophila. *Genetics,* 183**,** 1175-85.

KALMYKOVA, A. I., NURMINSKY, D. I., RYZHOV, D. V. & SHEVELYOV, Y. Y. 2005. Regulated chromatin domain comprising cluster of co-expressed genes in Drosophila melanogaster. *Nucleic Acids Research,* 33**,** 1435-1444.

KHIL, P. P., SMIRNOVA, N. A., ROMANIENKO, P. J. & CAMERINI-OTERO, R. D. 2004. The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nat Genet,* 36**,** 642-6.

KINSELLA, R. J., KÄHÄRI, A., HAIDER, S., ZAMORA, J., PROCTOR, G., SPUDICH, G., ALMEIDA-KING, J., STAINES, D., DERWENT, P., KERHORNOU, A., KERSEY, P. & FLICEK, P. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database,* 2011.

LERCHER, M. J., URRUTIA, A. O. & HURST, L. D. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet,* 31**,** 180-3.

LERCHER, M. J., URRUTIA, A. O. & HURST, L. D. 2003. Evidence that the human X chromosome is enriched for male-specific but not female-specific genes. *Mol Biol Evol,* 20**,** 1113-6.

MANK, J., HULTIN-ROSENBERG, L., WEBSTER, M. & ELLEGREN, H. 2008. The unique genomic properties of sex-biased genes: Insights from avian microarray data. *BMC Genomics,* 9**,** 148.

MANK, J. E., AXELSSON, E. & ELLEGREN, H. 2007a. Fast-X on the Z: rapid evolution of sex-linked genes in birds. *Genome Res,* 17**,** 618-24.

MANK, J. E., HULTIN-ROSENBERG, L., AXELSSON, E. & ELLEGREN, H. 2007b. Rapid evolution of female-biased, but not male-biased, genes expressed in the avian brain. *Mol Biol Evol,* 24**,** 2698-706.

MANK, J. E., NAM, K., BRUNSTROM, B. & ELLEGREN, H. 2010. Ontogenetic complexity of sexual dimorphism and sex-specific selection. *Mol Biol Evol,* 27**,** 1570-8.

MARINOTTI, O., CALVO, E., NGUYEN, Q. K., DISSANAYAKE, S., RIBEIRO, J. M. & JAMES, A. A. 2006. Genome-wide analysis of gene expression in adult Anopheles gambiae. *Insect Mol Biol,* 15**,** 1-12.

MEIKLEJOHN, C. D., PARSCH, J., RANZ, J. M. & HARTL, D. L. 2003. Rapid evolution of male-biased gene expression in Drosophila. *Proc Natl Acad Sci U S A,* 100**,** 9894-9.

MEISEL, R. P., MALONE, J. H. & CLARK, A. G. 2012. Disentangling the relationship between sex-biased gene expression and X-linkage. *Genome Res,* 22**,** 1255-65.

METTA, M., GUDAVALLI, R., GIBERT, J. M. & SCHLOTTERER, C. 2006. No accelerated rate of protein evolution in male-biased Drosophila pseudoobscura genes. *Genetics,* 174**,** 411-20.

MIKHAYLOVA, L. M. & NURMINSKY, D. I. 2011. Lack of global meiotic sex chromosome inactivation, and paucity of tissue-specific gene expression on the Drosophila X chromosome. *BMC Biol,* 9**,** 29.

MUELLER, J. L., MAHADEVAIAH, S. K., PARK, P. J., WARBURTON, P. E., PAGE, D. C. & TURNER, J. M. 2008. The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. *Nat Genet,* 40**,** 794-9.

NAURIN, S., HANSSON, B., HASSELQUIST, D., KIM, Y.-H. & BENSCH, S. 2011. The sex-biased brain: sexual dimorphism in gene expression in two species of songbirds. *BMC Genomics,* 12**,** 37.

PARISI, M., NUTTALL, R., EDWARDS, P., MINOR, J., NAIMAN, D., LU, J., DOCTOLERO, M., VAINER, M., CHAN, C., MALLEY, J., EASTMAN, S. & OLIVER, B. 2004. A survey of ovary-, testis-, and soma-biased gene expression in Drosophila melanogaster adults. *Genome Biol,* 5**,** R40.

PARISI, M., NUTTALL, R., NAIMAN, D., BOUFFARD, G., MALLEY, J., ANDREWS, J., EASTMAN, S. & OLIVER, B. 2003. Paucity of genes on the Drosophila X chromosome showing male-biased expression. *Science,* 299**,** 697-700.

PERRY, J. C., HARRISON, P. W. & MANK, J. E. 2014. The Ontogeny and Evolution of Sex-Biased Gene Expression in Drosophila melanogaster. *Molecular Biology and Evolution*.

PROSCHEL, M., ZHANG, Z. & PARSCH, J. 2006. Widespread adaptive evolution of Drosophila genes with sex-biased expression. *Genetics,* 174**,** 893-900.

PURMANN, A., TOEDLING, J., SCHUELER, M., CARNINCI, P., LEHRACH, H., HAYASHIZAKI, Y., HUBER, W. & SPERLING, S. 2007. Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality. Genomics, 89, 580-587.

RANZ, J. M., CASTILLO-DAVIS, C. I., MEIKLEJOHN, C. D. & HARTL, D. L. 2003. Sex-dependent gene expression and evolution of the Drosophila transcriptome. *Science,* 300**,** 1742-5.

REINIUS, B., SAETRE, P., LEONARD, J. A., BLEKHMAN, R., MERINO-MARTINEZ, R., GILAD, Y. & JAZIN, E. 2008. An evolutionarily conserved sexual signature in the primate brain. *PLoS Genet,* 4**,** e1000100.

REINKE, V., GIL, I. S., WARD, S. & KAZMER, K. 2004. Genome-wide germline-enriched and sex-biased expression profiles in Caenorhabditis elegans. *Development,* 131**,** 311-23.

RICE, W. R. 1984. Sex Chromosomes and the Evolution of Sexual Dimorphism. *Evolution,* 38**,** 735-742.

SMALL, C. M., CARNEY, G. E., MO, Q., VANNUCCI, M. & JONES, A. G. 2009. A microarray analysis of sex- and gonad-biased gene expression in the

zebrafish: evidence for masculinization of the transcriptome. *BMC Genomics,* 10**,** 579.

SPELLMAN, P. T. & RUBIN, G. M. 2002. Evidence for large domains of similarly expressed genes in the Drosophila genome. *J Biol,* 1**,** 5.

STURGILL, D., ZHANG, Y., PARISI, M. & OLIVER, B. 2007. Demasculinization of X chromosomes in the Drosophila genus. *Nature,* 450**,** 238-41.

SWANSON, W. J. & VACQUIER, V. D. 2002. The rapid evolution of reproductive proteins. *Nat Rev Genet,* 3**,** 137-44.

VERSTEEG, R., VAN SCHAIK, B. D., VAN BATENBURG, M. F., ROOS, M., MONAJEMI, R., CARON, H., BUSSEMAKER, H. J. & VAN KAMPEN, A. H. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res,* 13**,** 1998-2004.

WANG, P. J., MCCARREY, J. R., YANG, F. & PAGE, D. C. 2001. An abundance of X-linked genes expressed in spermatogonia. *Nat Genet,* 27**,** 422-6.

YANAI, I., BENJAMIN, H., SHMOISH, M., CHALIFA-CASPI, V., SHKLAR, M., OPHIR, R., BAR-EVEN, A., HORN-SABAN, S., SAFRAN, M., DOMANY, E., LANCET, D. & SHMUELI, O. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics,* 21**,** 650-9.

YANG, X., SCHADT, E. E., WANG, S., WANG, H., ARNOLD, A. P., INGRAM-DRAKE, L., DRAKE, T. A. & LUSIS, A. J. 2006. Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res,* 16**,** 995-1004.

ZHANG, Y., STURGILL, D., PARISI, M., KUMAR, S. & OLIVER, B. 2007. Constraint and turnover in sex-biased gene expression in the genus Drosophila. *Nature,* 450**,** 233-7.

ZHANG, Y. E., VIBRANOVSKI, M. D., KRINSKY, B. H. & LONG, M. 2010. Age-dependent chromosomal distribution of male-biased genes in Drosophila. *Genome Res,* 20**,** 1526-33.

ZHANG, Z., HAMBUCH, T. M. & PARSCH, J. 2004. Molecular evolution of sex-biased genes in Drosophila. *Mol Biol Evol,* 21**,** 2130-9.

ZHANG, Z. & PARSCH, J. 2005. Positive correlation between evolutionary rate and recombination rate in Drosophila genes with male-biased expression. *Mol Biol Evol,* 22**,** 1945-7.

2.7 Tables and figures

Table 2-2. Summary of Female:Male expression ratio (all tissues average)

Species	Mean	Std. err.	Male biased	Female biased	χ^2 test p-value
Gorilla	-0.114078	0.009106	6347	6367	0.8592
Human	-0.308330	0.009489	4651	7949	< 2.2e-16
Macaque	-0.051459	0.008606	6755	5846	5.601e-16
Bonobo	-0.092863	0.009256	5549	7162	< 2.2e-16
Orangutan	-0.107872	0.010805	6115	6252	0.218
Chimp	-0.101687	0.010278	5792	6958	< 2.2e-16

Table 2-3. Correlation between tau and absolute Female:Male expression ratio (log_2)

Species	Pearson correlation coefficient	p-value
Gorilla	0.4371604	< 2.2e-16
Human	0.3767478	< 2.2e-16
Macaque	0.3849789	< 2.2e-16
Bonobo	0.4068094	< 2.2e-16
Orangutan	0.3606426	< 2.2e-16
Chimp	0.3639032	< 2.2e-16

Table 2-4. Correlation between average expression ratio (log_2) and absolute Female:Male expression ratio (log_2)

Species	Pearson correlation coefficient	p-value
Gorilla	-0.189669	< 2.2e-16
Human	-0.1420363	< 2.2e-16
Macaque	-0.20696	< 2.2e-16
Bonobo	-0.2194935	< 2.2e-16
Orangutan	-0.1484856	< 2.2e-16
Chimp	-0.2002176	< 2.2e-16

Table 2-5. Correlation between dN/dS and Female:Male expression ratio (log_2)

Species	Pearson correlation coefficient	p-value
Gorilla	-0.01343023	0.1797
Human	-0.02389866	0.01741
Macaque	-0.01723596	0.0861
Bonobo	NA	NA
Orangutan	-0.006124627	0.5457
Chimp	0.006756152	0.4994

Table 2-6. Difference in distribution of sex biased gene expression among chromosomes.

Species	Kruskal-Wallis chi-squared	df	p-value
Gorilla	34.6729	23	0.05603
Human	86.4129	22	1.387e-09
Macaque	225.437	20	< 2.2e-16
Orangutan	211.9557	23	< 2.2e-16
Chimp	41.1331	23	0.01143

Table 2-7. Correlation between testis genes and sex biased expressed genes

Species	Pearson correlation coefficient	p-value
Gorilla	0.06804319	4.614e-13
Human	0.1451405	< 2.2e-16
Macaque	0.1745345	< 2.2e-16
Bonobo[a]	na	na
Orangutan[b]	na	na
Chimp	0.1794017	< 2.2e-16

[a]Chromosome assembly information not available at the stage analysed.
[b]Testis gene expression data not available

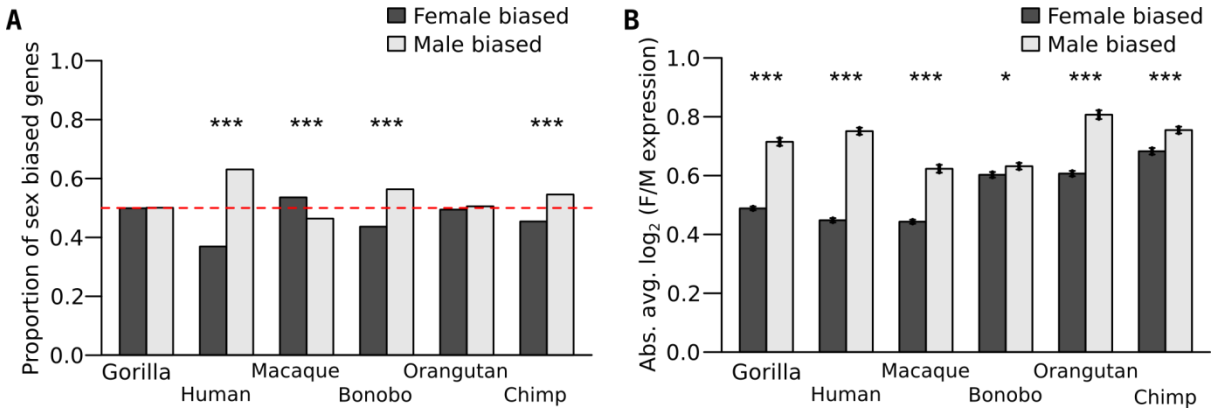


Figure 2-1.Barplots showing differences in distributions of the average degree of sex-biased expression within species. (A) Proportion of genes with an average female-biased expression pattern (log2 (F/M expression) > 0) and male-biased (log2 (F/M expression) <0) genes in six primate species. Significant differences from the expected proportion as assessed by Chi-squared test are denoted with asterisks (*** p< 0.001). Dashed line represents the expected proportion of female-biased genes (50%). (B) Mean strength of the average degree of sex-biased expression in female-biased (dark grey) and male-biased (light grey) genes per primate. Whiskers indicate the standard error of the mean. Significant differences between female and male are marked with asterisks. (T-test, * p< 0.5, *** p<0.001).

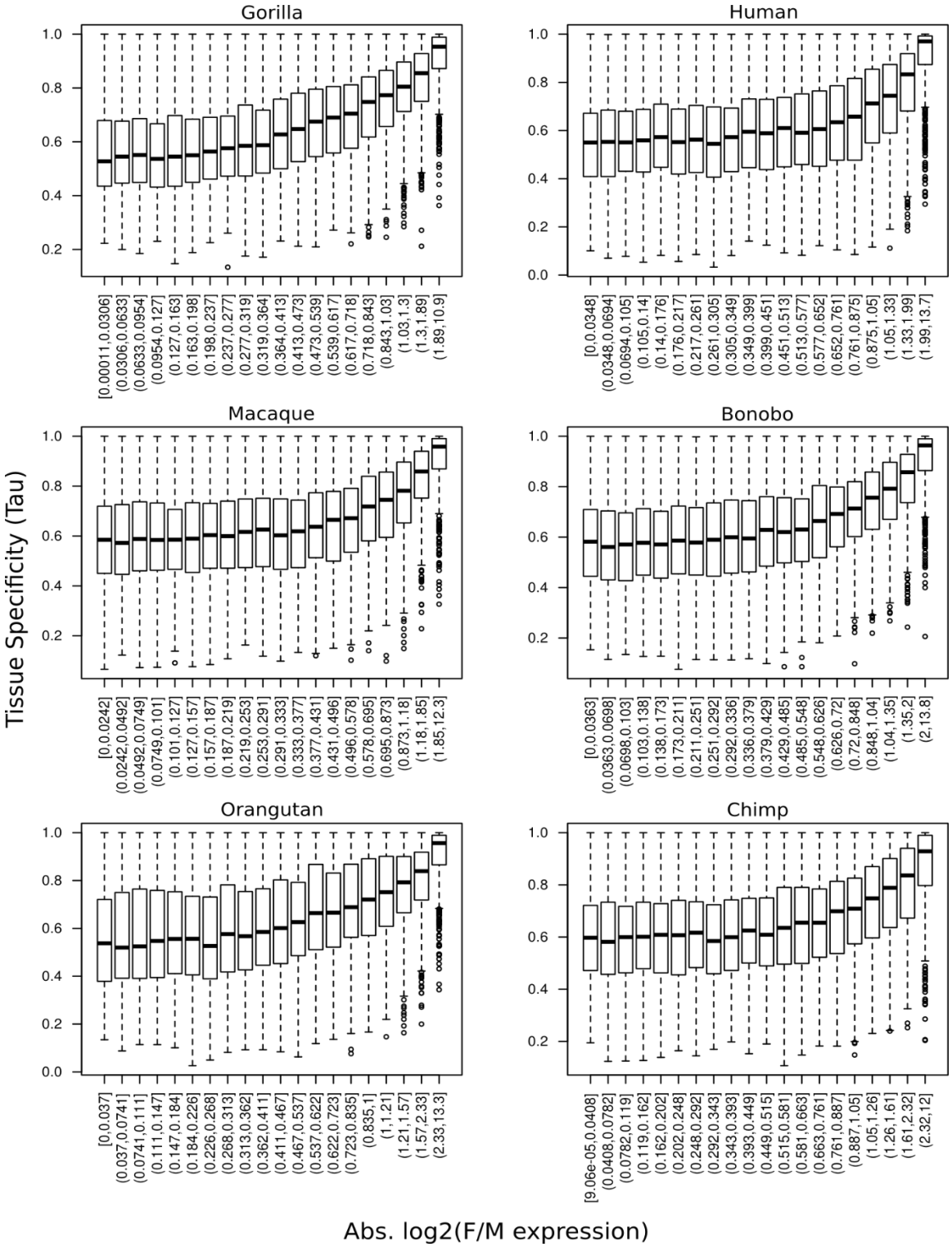


Figure 2-2. Boxplot showing trend between the absolute sex-biased expression averaged across tissues (X-axis) and tissue specificity (Y-axis) per primate. Boxes denote interquartile ranges, lines denote medians, and whiskers denote 1.5 times the interquartile range. Each box in the boxplot was constructed by dividing the genes into vigintiles according to the strength of the sex-biased in their gene expression.

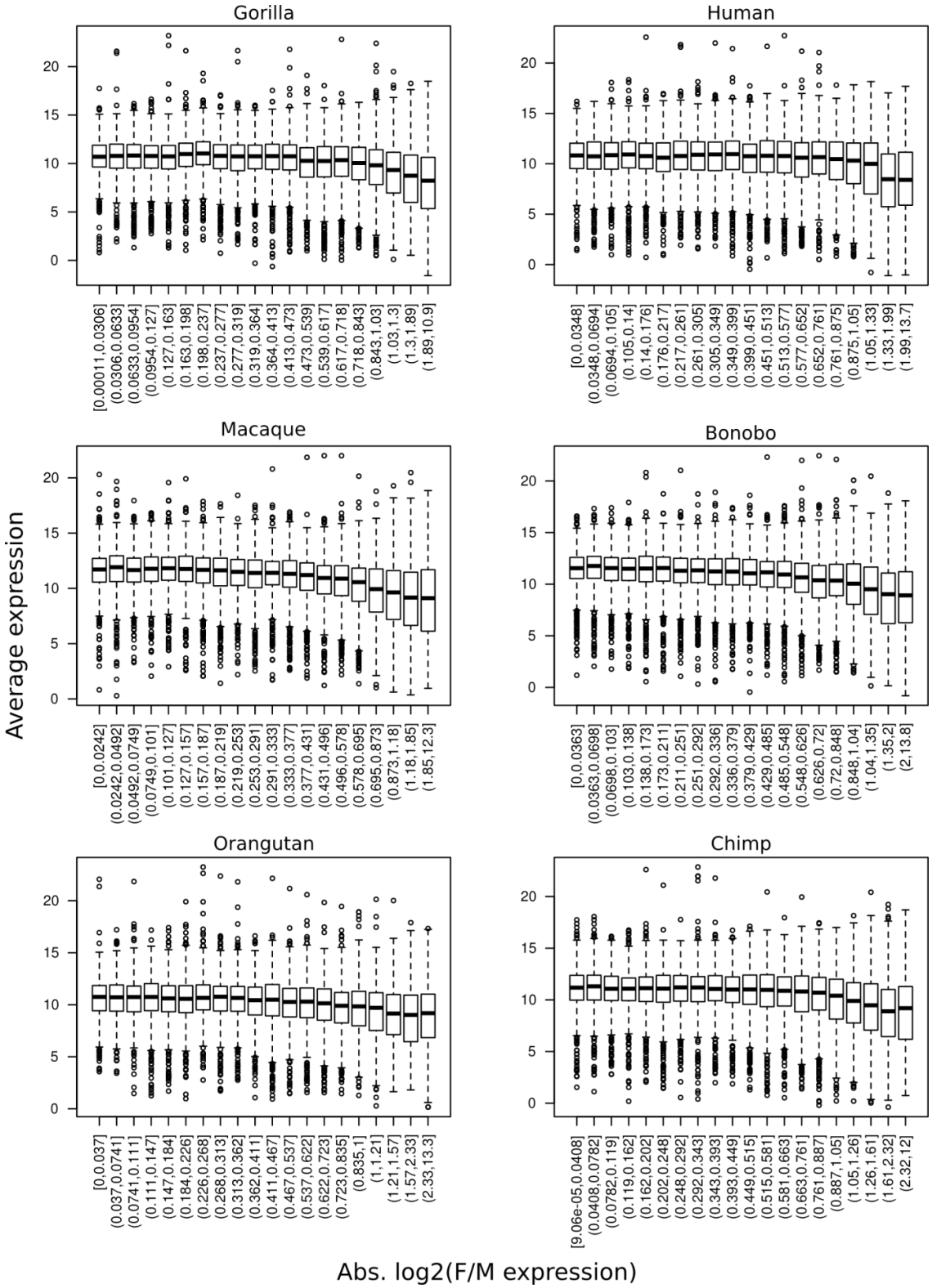


Figure 2-3. Boxplot showing trend between the absolute sex-biased expression averaged across tissues (X-axis) and the gene's average gene expression (Y-axis) in the analysed tissues for 6 primate species. Boxes denote interquartile ranges, lines denote medians, and whiskers denote 1.5 times the interquartile range. Each box in the boxplot was constructed by dividing the genes into vigintiles according to the strength of the sex-biased in their gene expression.

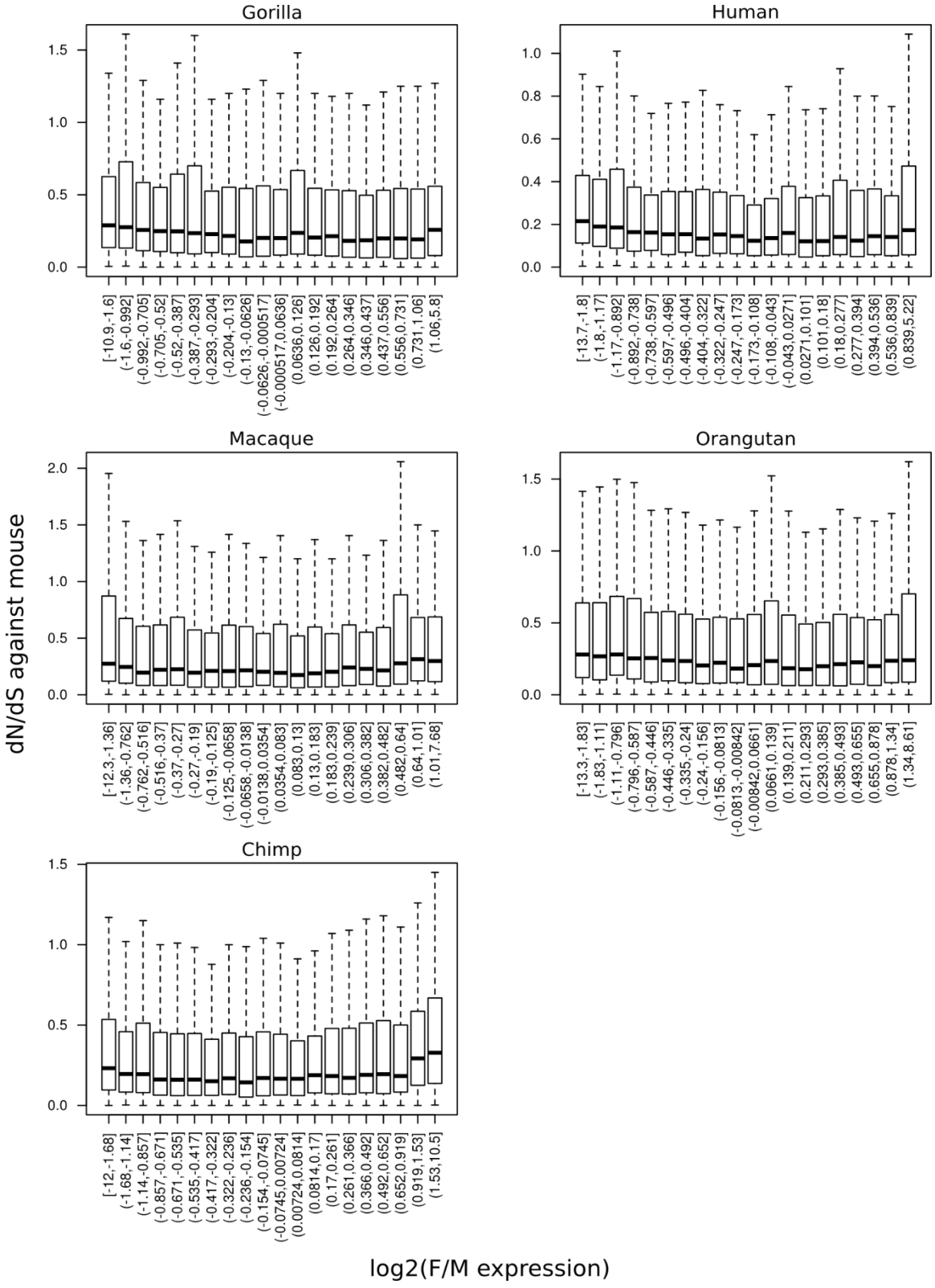


Figure 2-4. Boxplot showing trend between the sex-bias in gene expression averaged across tissues (X-axis) and the gene'sdN/dS computed in comparison against the mouse (Y-axis) for 5 primate species with complete genome assemblies available. Boxes denote interquartile ranges, lines denote medians, and whiskers denote 1.5 times the interquartile range. Each box in the boxplot was constructed by dividing

the genes into vigintiles according to the strength of the sex-biased in their gene expression.

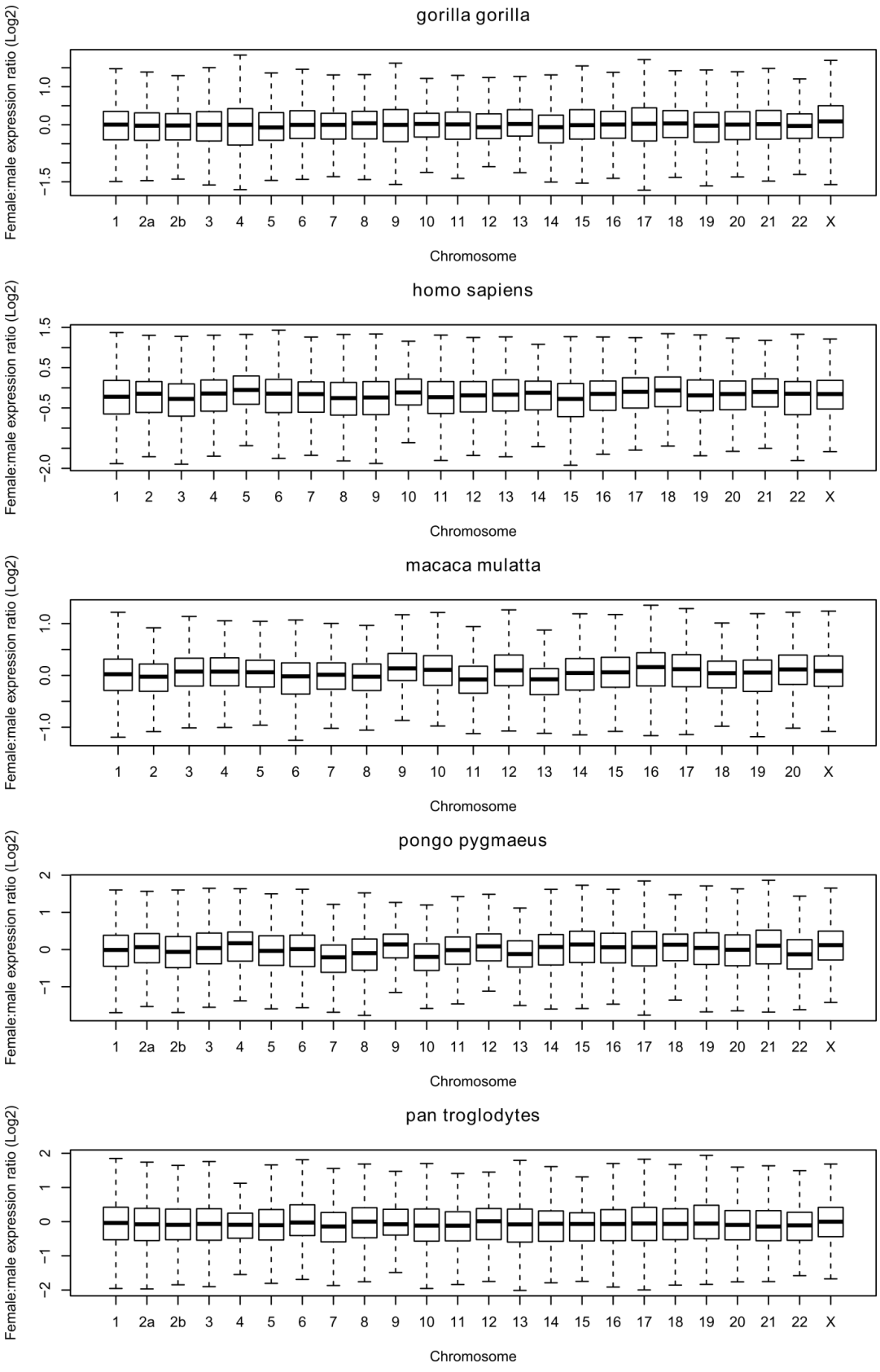
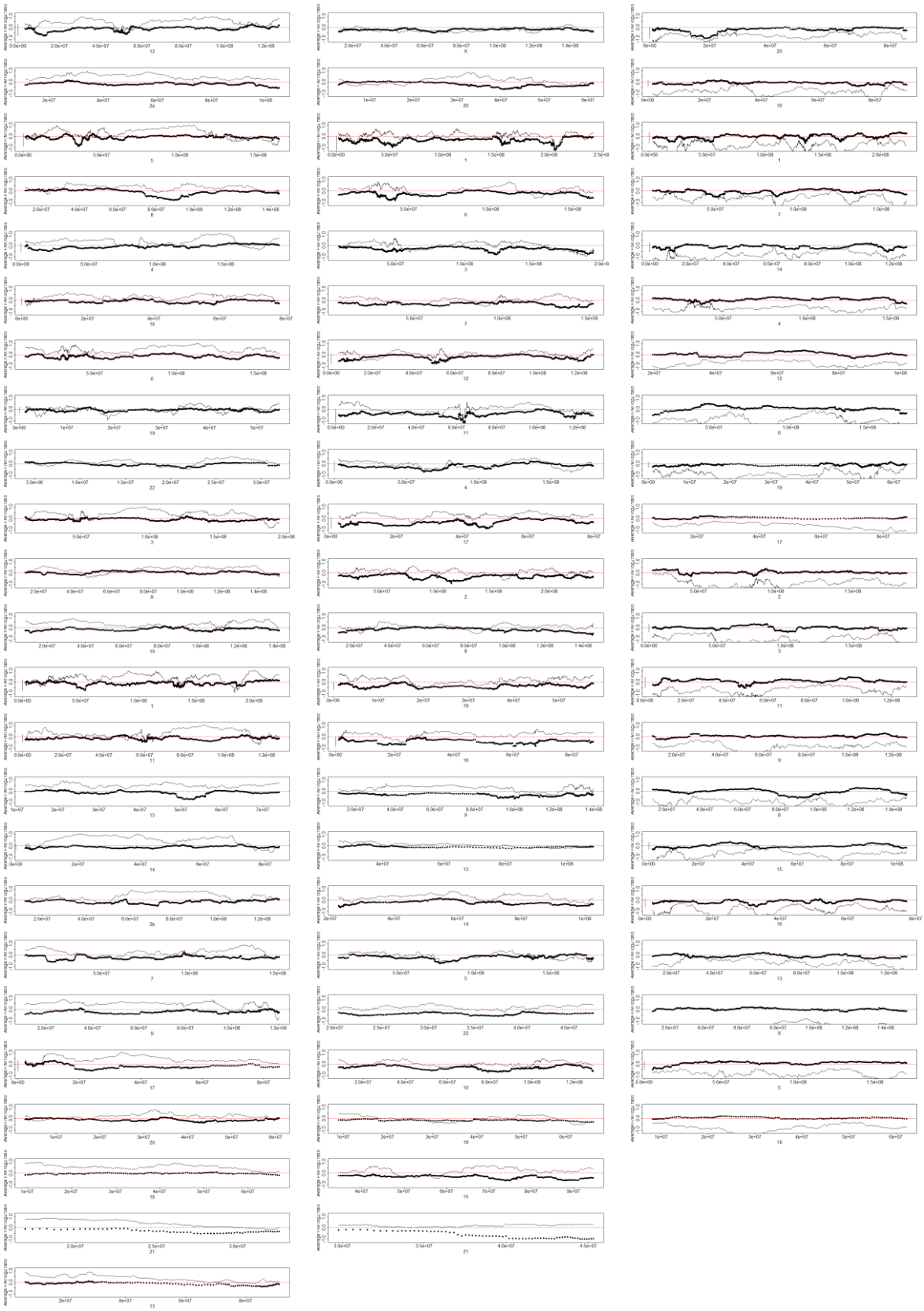
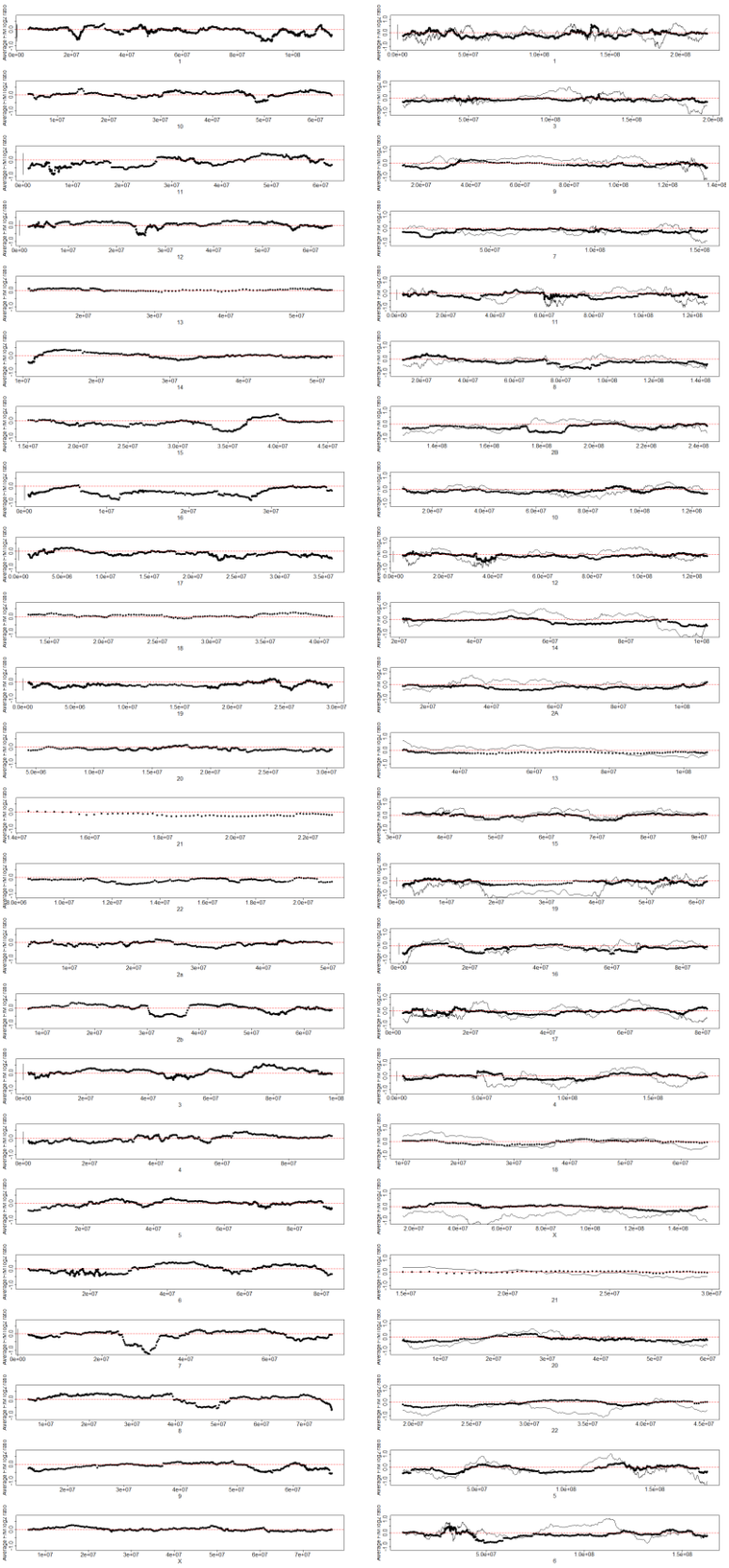


Figure 2-5. Difference in distribution of sex-biased gene expression among chromosomes. Boxplots showing the distributions of $\log_2$(female to male expression ratio) for the genes in each chromosome. Boxes denote interquartile ranges, mid lines the median and whiskers 1.5 times the interquartile range.



Gorilla Human Macaque



Orangutan Chimp

Figure 2-6. Sliding window dotplots of female/male-biased gene clustering (bold dotted lines) and testis gene clustering (thin lines) analyses, mapped by position on each chromosome across species. Each dot represents the average sex-biased gene expression ratio in a sliding window of 50 genes. The x-axis shows the mid-position in base pairs of the 50-gene window.

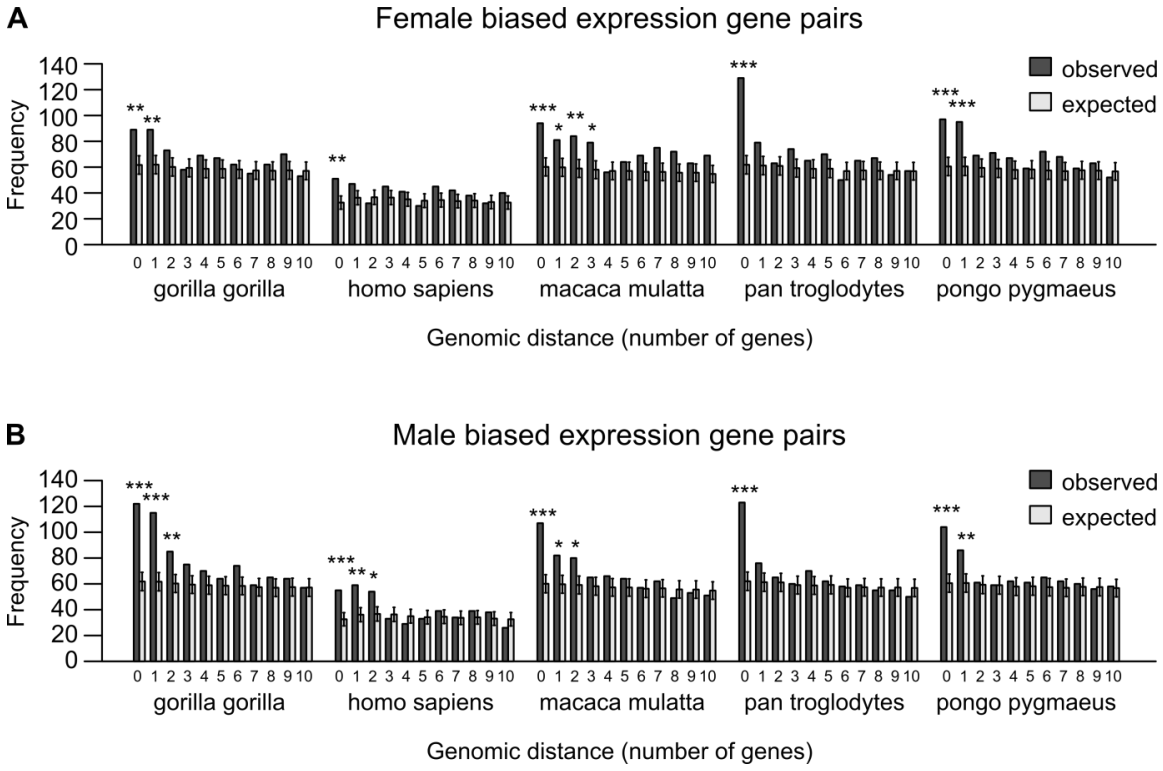


Figure 2-7. Neighbouring genes are more likely to have highly sex-biased expression. We counted gene pairs with similar female:male expression ratios separated by an increasing number of genes and compared the observed counts to those expected from a Monte Carlo simulation based on 10,000 randomisations. In order to consider a gene pair as having similar sex-biased expression, both genes had to be included within the top 10% of the distribution for female:male, or male:female, expression ratio. Barplots show the observed and expected number of neighbouring gene pairs with (A) female-biased, or (B) male-biased expression (y-axis) separated by an increasing number of genes (x-axis) for each primate. Error bars denote the standard error of the 10,000 random samples (Bonferroni adjusted p-values *$p < 0.05$, **$p < 0.01$, *** $p< 0.001$).

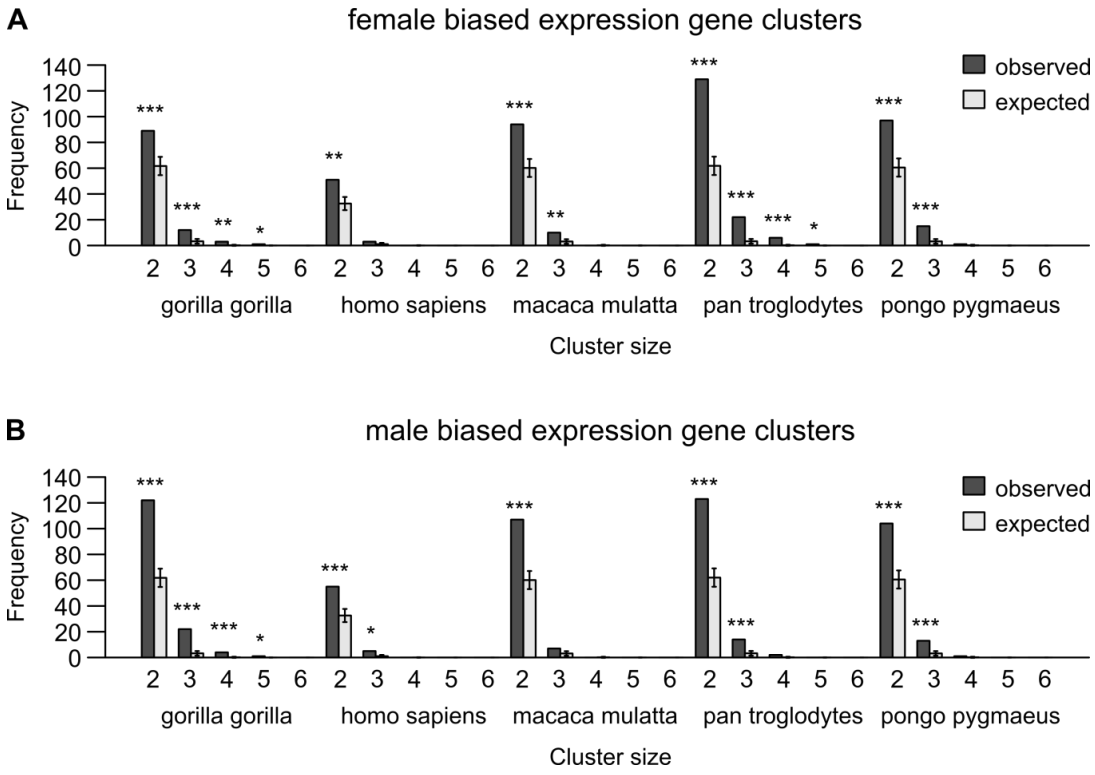


Figure 2-8. Enrichment of clusters with the highest sex-biased expression. We selected the top 10% of genes with the highest male-biased expression and the top 10% with the highest female-biased expression. We count the numbers and size of gene clusters within the highest biased expression and compared it to the expected number of equally sized clusters obtained through 10,000 randomizations. Barplots show the observed and expected number of A) female- and B) male-biased expression gene clusters in each of the five studied primate species. Error bars denote the standard error of the 10,000 random samples (Bonferroni adjusted p values *p < 0.05, **p < 0.01, *** p< 0.001).

2.8 Supplementary tables and figures

Table S2-8.Gene expression data availability of each tissue in six primate species.

Lineage	Brain	Cerebellum	Heart	Kidney	Liver	Testis
Gorilla	√	√	√	√	√	√
Human	√	√	√	√	NA	√
Macaque	√	√	√	√	√	√
Bonobo	√	√	√	√	√	√
Orangutan	√	NA	√	√	√	NA
Chimp	√	√	√	√	√	√

Table S2-9. ANOVA test for sex-biased expression distribution between tissues in six primate species.

Lineage	F	p value
Gorilla	23296	<2e-16
Human	95805	<2e-16
Macaque	44149	<2e-16
Bonobo	29549	<2e-16
Orangutan	34351	<2e-16
Chimp	52186	<2e-16

Table S2-10. Summary of female:male expression ratio (Brain).

Species	Female (%)	Male (%)	Chi^2	p-value
Gorilla	39.62	60.38	513.8516	9.21E-114
Human	55.21	44.79	131.6255	1.81E-30
Macaque	65.80	34.20	1185.109	1.05E-259
Bonobo	41.93	58.07	316.1968	9.76E-71
Orangutan	63.07	36.93	786.6957	4.21E-173
Chimp	64.06	35.94	957.1465	3.71E-210

Table S2-11. Summary of female:male expression ratio (Cerebellum).

Species	Female (%)	Male (%)	Chi^2	p-value
Gorilla	64.60	35.40	988.7027	5.13E-217
Human	38.69	61.31	600.4873	1.31E-132
Macaque	57.55	42.45	262.0093	6.26E-59
Bonobo	26.90	73.10	2504.239	0
Orangutan	45.04	54.96	114.1775	1.19E-26
Chimp	64.60	35.40	988.7027	5.13E-217

Table S2-12. Summary of female:male expression ratio (Heart).

Species	Female (%)	Male (%)	Chi^2	p-value
Gorilla	73.63	26.37	2416.650	0
Human	13.41	86.59	6141.987	0
Macaque	23.10	76.90	3225.346	0
Bonobo	61.91	38.09	634.6372	4.90E-140
Orangutan	86.22	13.78	5459.652	0
Chimp	53.30	46.70	48.55955	3.20E-12

Table S2-13. Summary of female:male expression ratio (Kidney).

Species	Female (%)	Male (%)	Chi^2	p-value
Gorilla	79.10	20.90	3968.487	0
Human	78.54	21.46	3868.90	0
Macaque	84.78	15.22	5504.605	0
Bonobo	64.52	35.48	994.870	2.34E-218
Orangutan	20.21	79.79	4053.928	0
Chimp	46.25	53.75	67.24879	2.39E-16

Table S2-14. Summary of female:male expression ratio (Liver).

Species	Female (%)	Male (%)	Chi^2	p-value
Gorilla	39.00	61.00	558.5756	1.72E-123
Human	63.30	36.70	765.6576	1.58E-168
Macaque	45.98	54.02	72.62851	1.57E-17
Bonobo	43.39	56.61	190.6647	2.28E-43
Orangutan	63.15	36.85	753.0334	8.79E-166
Chimp	39.00	61.00	558.5756	1.72E-123

Table S2-15. Stats of correlation between average expression and female:male expression ratio (log2) in each tissue.

	Brain		Cerebellum		Heart		Kidney		Liver	
	cor*	**p value**	**cor**	**p value**	**cor**	**p value**	**cor**	**p value**	**cor**	**p value**
Gorilla	-0.017095	0.06168	0.224122	< 2.2e-16	0.186530	< 2.2e-16	-0.138356	< 2.2e-16	0.119651	< 2.2e-16
Human	0.066614	2.096e-13	0.073300	1.776e-15	-0.085224	< 2.2e-16	0.019494	0.03394	NA	NA
Macaque	-0.095587	< 2.2e-16	-0.007258	0.4363	0.071306	4.53e-14	0.000994	0.9156	-0.192292	< 2.2e-16
Bonobo	0.091812	< 2.2e-16	-0.082115	< 2.2e-16	0.244866	< 2.2e-16	0.080440	< 2.2e-16	0.229201	< 2.2e-16
Orangutan	-0.073288	3.118e-15	NA	NA	-0.078910	7.412e-16	-0.132229	< 2.2e-16	0.064203	1.777e-11
Chimp	-0.263510	< 2.2e-16	-0.239679	< 2.2e-16	-0.140390	< 2.2e-16	-0.044701	1.016e-06	0.023792	0.01297

* Pearson correlation coefficient

Table S2-16. Stats of correlation between tissue specificity and female:male expression ratio (log2) in each tissue.

	Brain		Cerebellum		Heart		Kidney		Liver	
	cor*	**p value**	**cor**	**p value**	**cor**	**p value**	**cor**	**p value**	**cor**	**p value**
Gorilla	-0.034958	0.0001323	-0.041727	6.754e-06	-0.020137	0.03599	0.100925	< 2.2e-16	-0.048858	1.441e-07
Human	-0.072474	1.348e-15	-0.036062	9.207e-05	0.007213	0.4398	-0.116069	< 2.2e-16	NA	NA
Macaque	0.020450	0.02583	0.018978	0.04176	-0.095128	< 2.2e-16	0.001881	0.8411	0.059550	5.564e-10
Bonobo	0.017819	0.0496	-0.009682	0.2938	-0.085831	< 2.2e-16	0.074677	4.441e-16	-0.054133	8.691e-09
Orangutan	-0.052003	2.245e-08	NA	NA	-0.052104	1.04e-07	0.104957	< 2.2e-16	-0.044312	3.52e-06
Chimp	0.002788	0.759	0.011667	0.2083	0.038065	5.767e-05	0.004965	0.5873	-0.153991	< 2.2e-16

* Pearson correlation coefficient

Table S2-17. Stats of correlation between dN/dS and female:male expression ratio (log_2) in each tissue.

	Brain		Cerebellum		Heart		Kidney		Liver	
	cor*	**p value**	**cor**	**p value**	**cor**	**p value**	**cor**	**p value**	**cor**	**p value**
Gorilla	0.006395	0.535	-0.024336	0.01986	-0.004118	0.7033	-0.015189	0.1443	-0.003321	0.7513
Human	-0.036076	0.0004225	0.005697	0.5835	0.014994	0.1547	-0.024884	0.01619	NA	NA
Macaque	0.017460	0.09068	-0.012420	0.2362	-0.020322	0.05636	-0.013309	0.2068	-0.006322	0.5588
Bonobo	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Orangutan	0.005494	0.5997	NA	NA	0.025600	0.02044	0.030224	0.00412	-0.010074	0.3498
Chimp	0.020141	0.04926	0.016439	0.1153	0.026778	0.01213	0.020625	0.04546	-0.003319	0.7582

* Pearson correlation coefficient

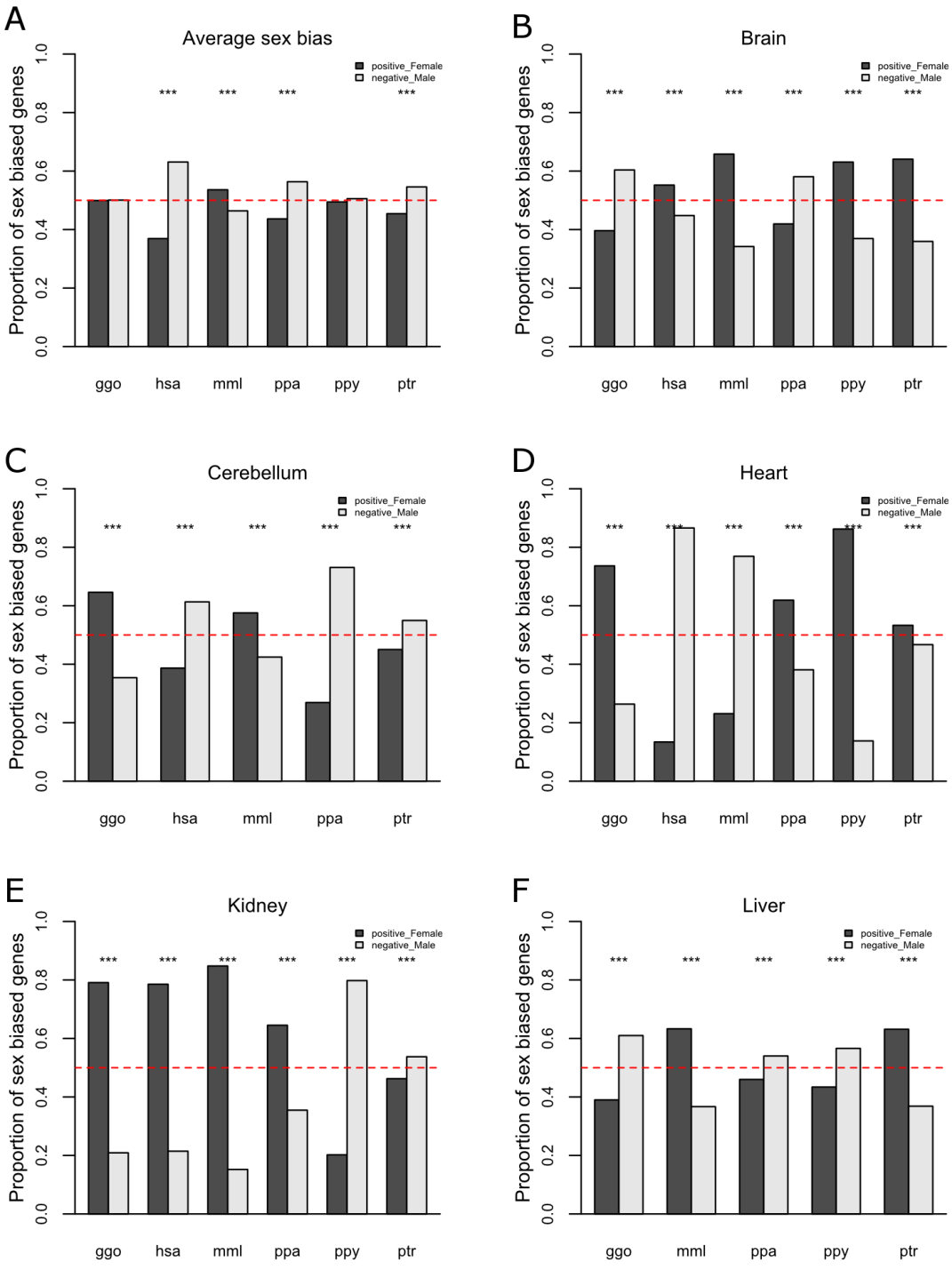


Figure S2-9. Barplots showing differences in distribution of the average expression sex bias within species. Proportion of genes with an average female biased expression pattern (log_2 (F/M expression) > 0) and male biased (log_2 (F/M expression) <0) genes in six primate species. A) averaged across tissues, B-F) each tissue. Significant differences from the expected proportion as assessed by Chi-squared test are denoted with asterisks (*** p< 0.001). Dashed line represents the expected proportion of female biased genes (50%).

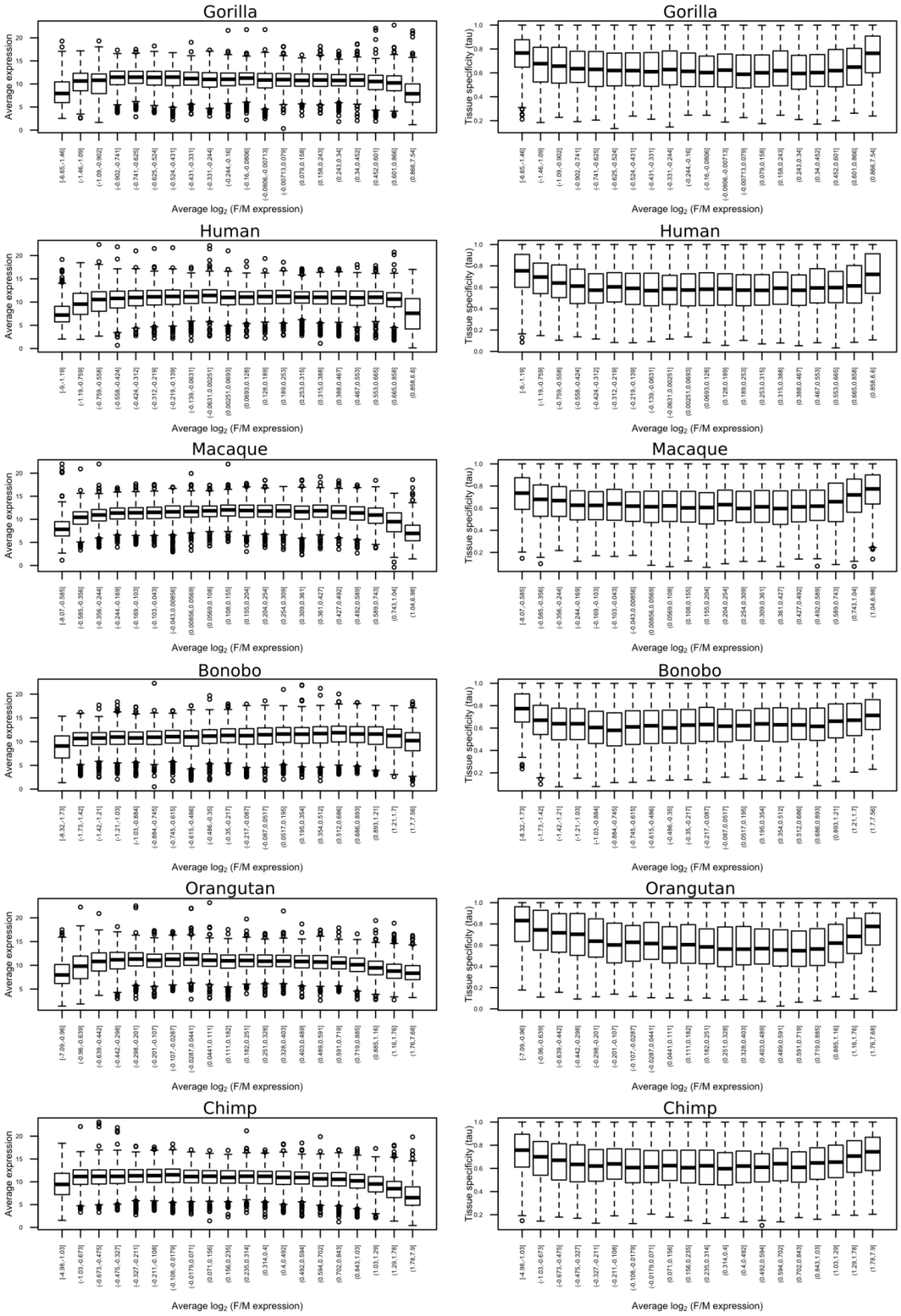


Figure S2-10. Boxplot showing trend between the absolute gene expression sex bias in the brain in the X-axis and average expression or tissue specificity (Chan, et al.) in the Y-axis

per primate species. Boxes denote interquartile ranges, lines denote medians, and whiskers denote 1.5 times the interquartile range. Each box in the boxplot was constructed by dividing the genes into vigintiles according to the strength of the sex biased in their gene expression.

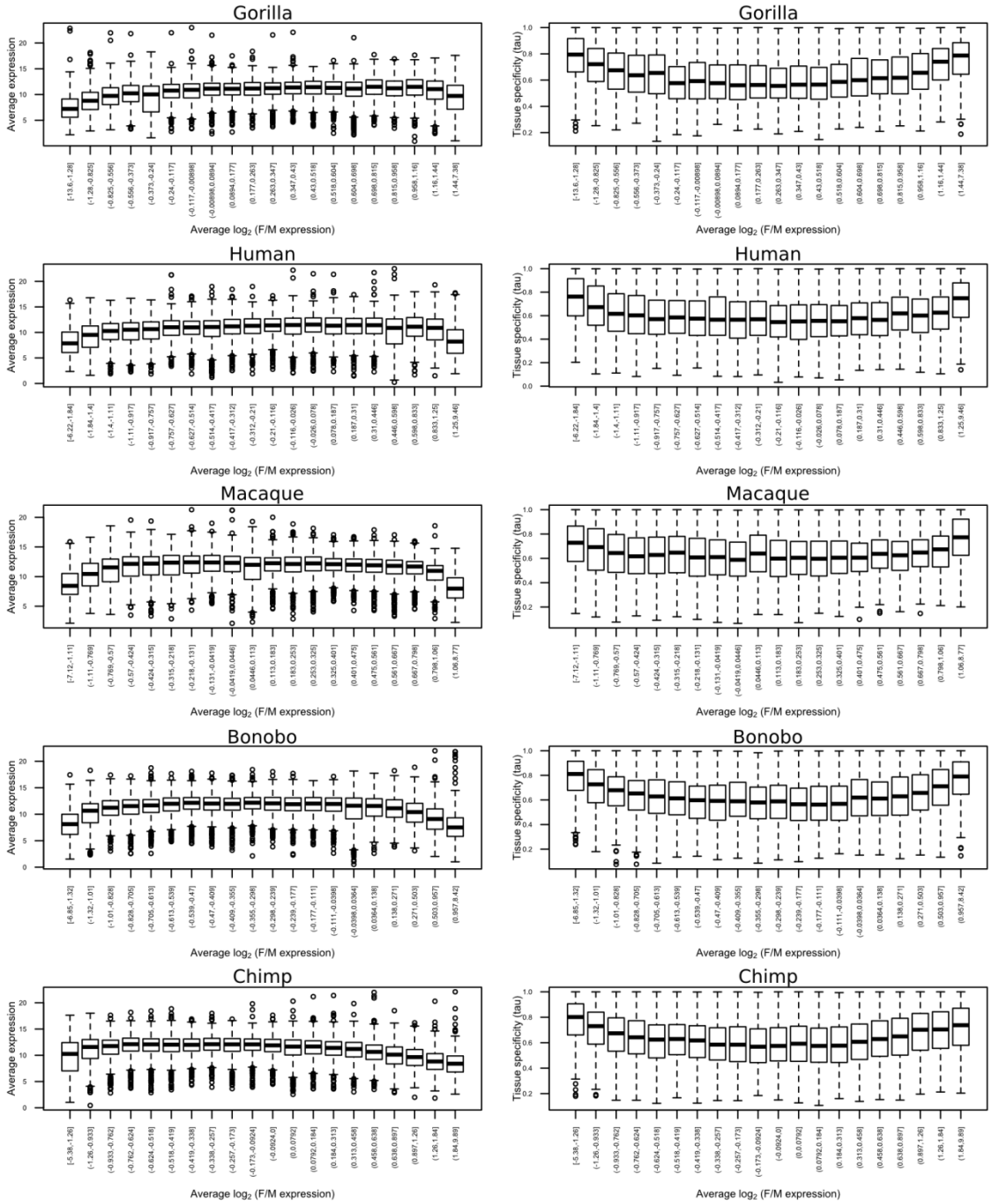


Figure S2-11. Boxplot showing trend between the absolute gene expression sex bias in the cerebellum in the X-axis and average expression or tissue specificity (Chan, et al.) in the Y-

axis per primate species. Boxes denote interquartile ranges, lines denote medians, and whiskers denote 1.5 times the interquartile range. Each box in the boxplot was constructed by dividing the genes into vigintiles according to the strength of the sex biased in their gene expression.

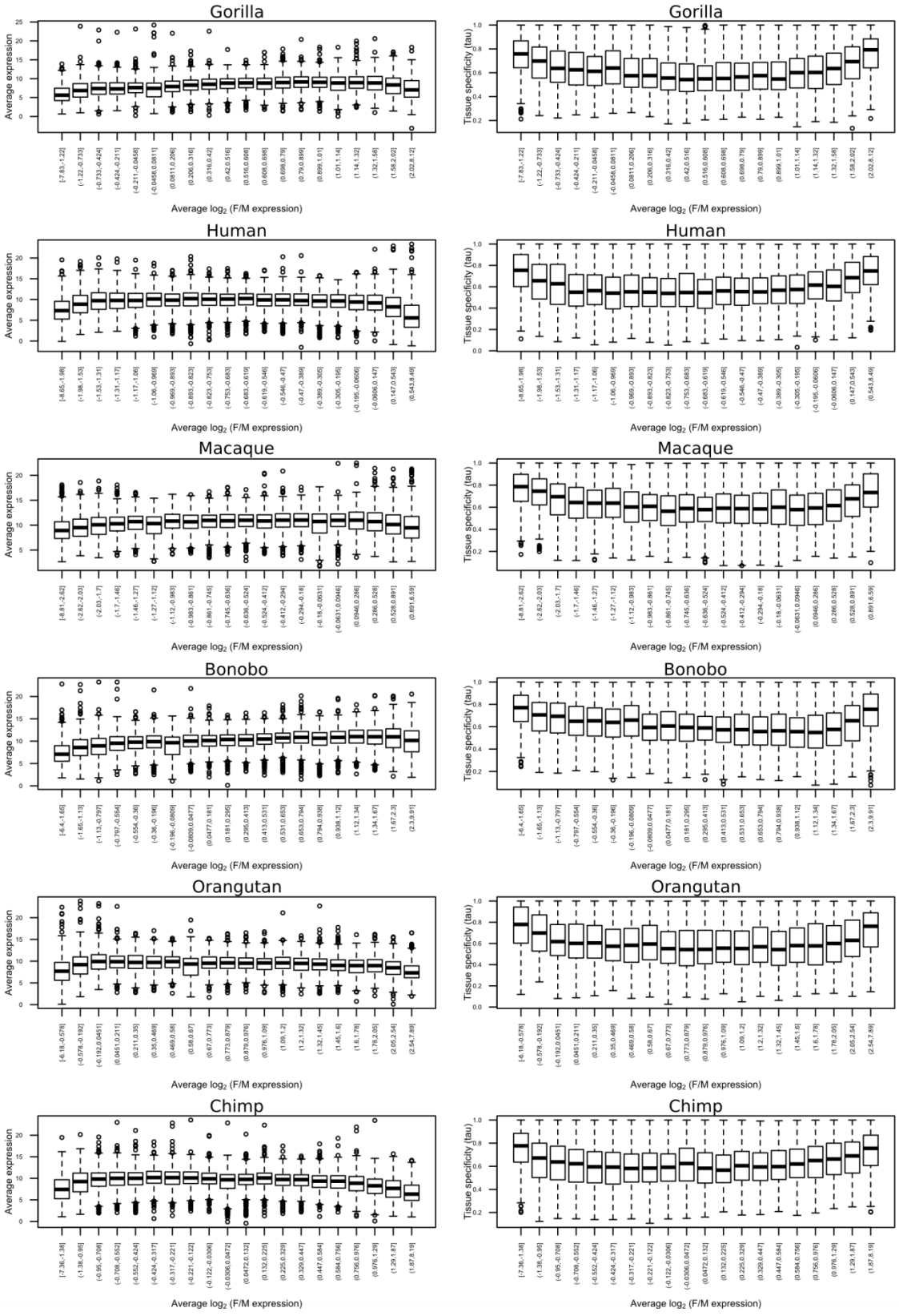


Figure S2-12. Boxplot showing trend between the absolute gene expression sex bias in the heart in the X-axis and average expression or tissue specificity (Chan, et al.) in the Y-axis

per primate species. Boxes denote interquartile ranges, lines denote medians, and whiskers denote 1.5 times the interquartile range. Each box in the boxplot was constructed by dividing the genes into vigintiles according to the strength of the sex biased in their gene expression.

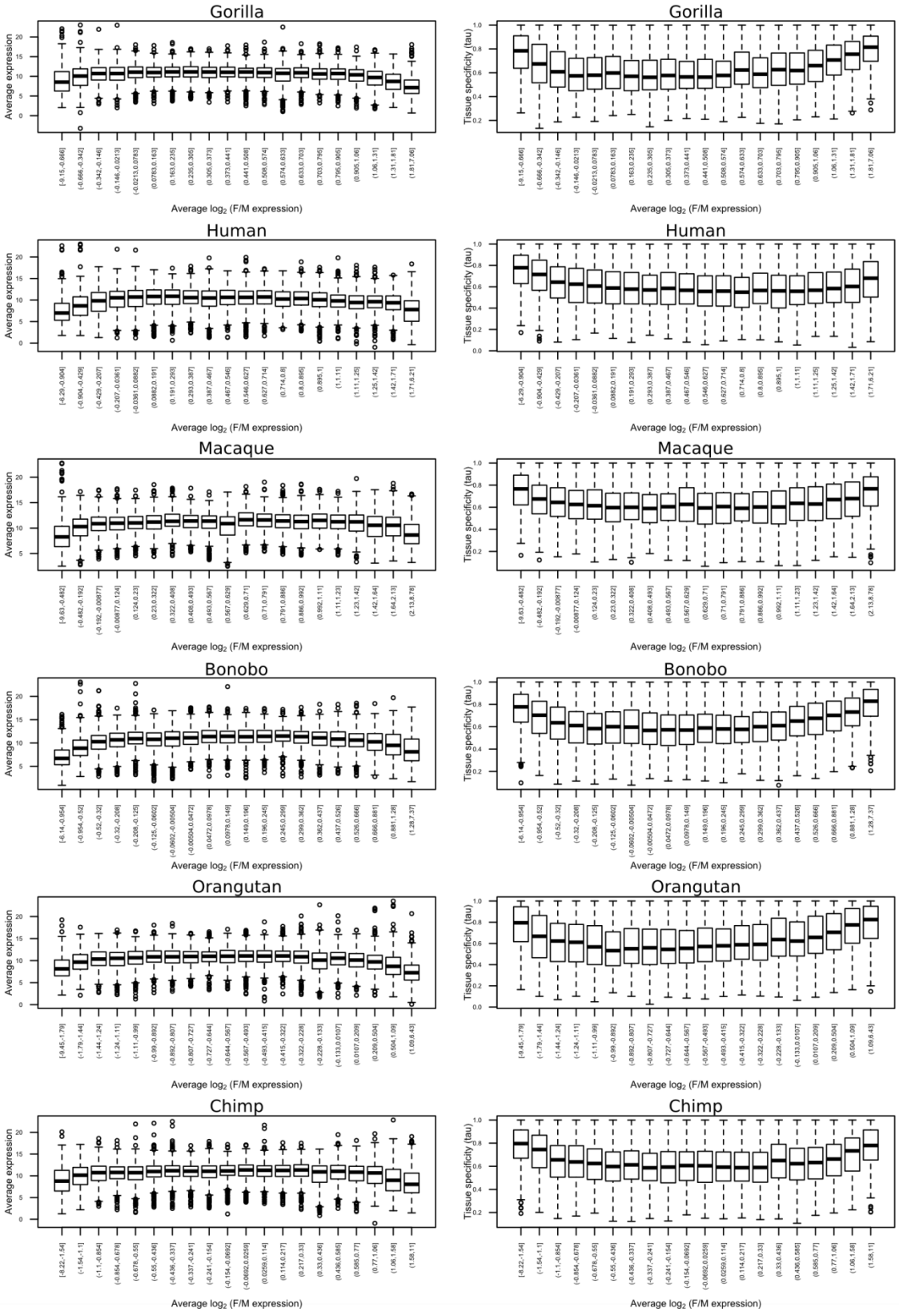


Figure S2-13. Boxplot showing trend between the absolute gene expression sex bias in the kidney in the X-axis and average expression or tissue specificity (Chan, et al.) in the Y-axis

per primate species. Boxes denote interquartile ranges, lines denote medians, and whiskers denote 1.5 times the interquartile range. Each box in the boxplot was constructed by dividing the genes into vigintiles according to the strength of the sex biased in their gene expression.

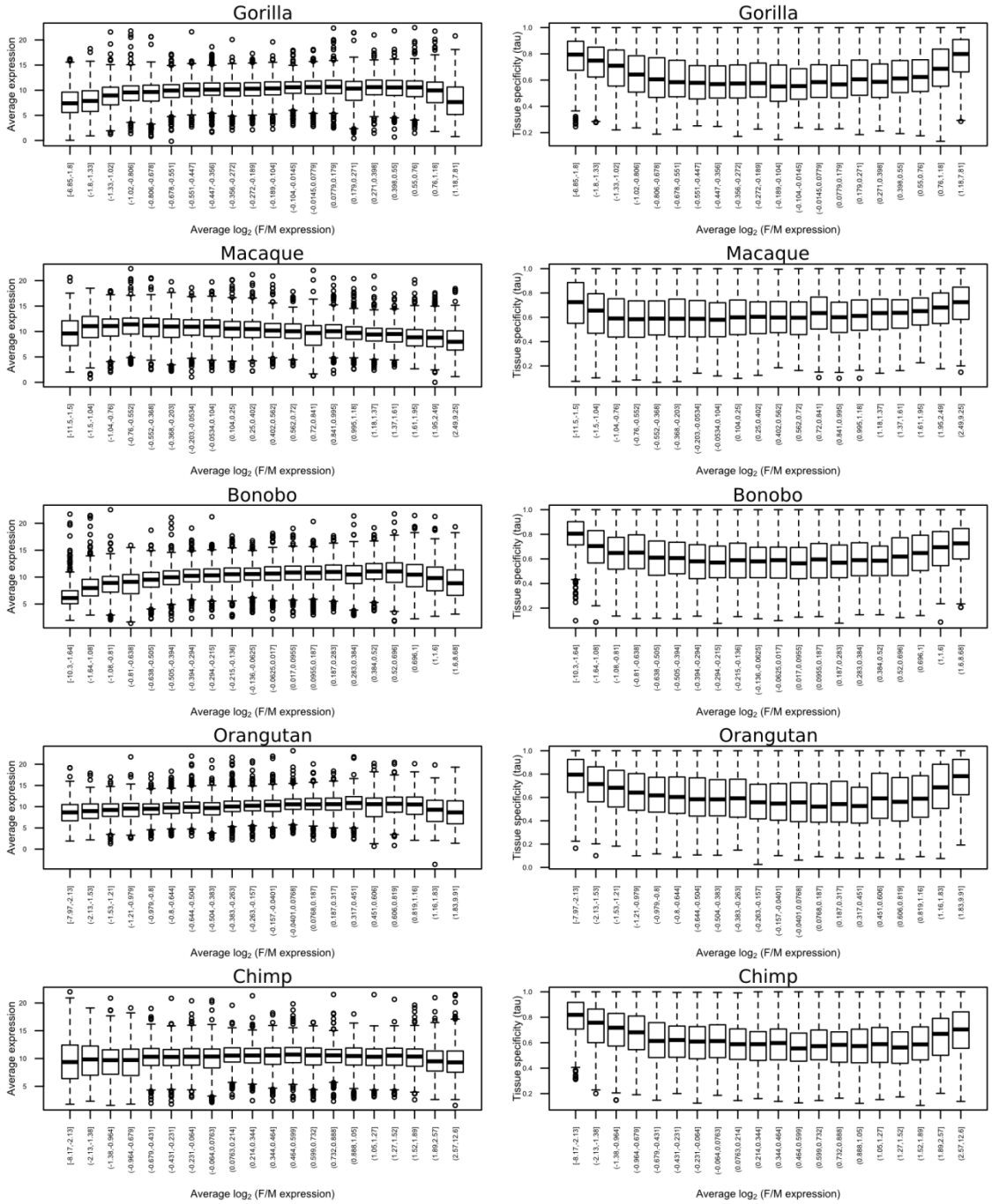


Figure S2-14. Boxplot showing trend between the absolute gene expression sex bias in the liver in the X-axis and average expression or tissue specificity (Chan, et al.) in the Y-axis

per primate species. Boxes denote interquartile ranges, lines denote medians, and whiskers denote 1.5 times the interquartile range. Each box in the boxplot was constructed by dividing the genes into vigintiles according to the strength of the sex biased in their gene expression.

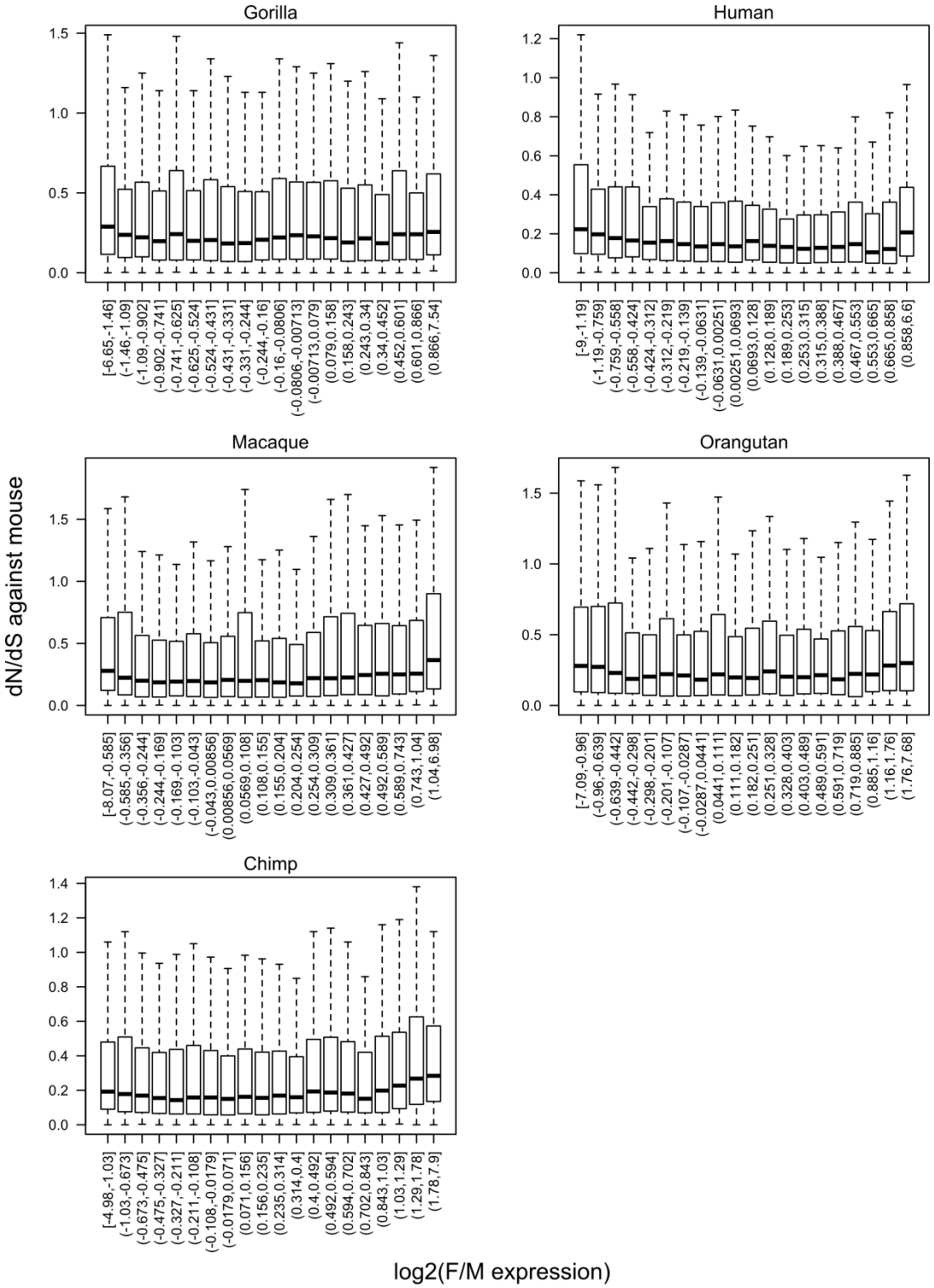


Figure S2-15. Boxplot showing trend between the sex bias of the expression of each gene in brain (X-axis) and its dN/dS computed comparing against mouse (Y-axis) for 5 primate species with complete genome assembly. Boxes denote interquartile ranges, lines denote medians, and whiskers denote 1.5 times the interquartile range. Each box in the boxplot was constructed by dividing the genes into 20 quantiles according to the strength of the sex biased in their gene expression.

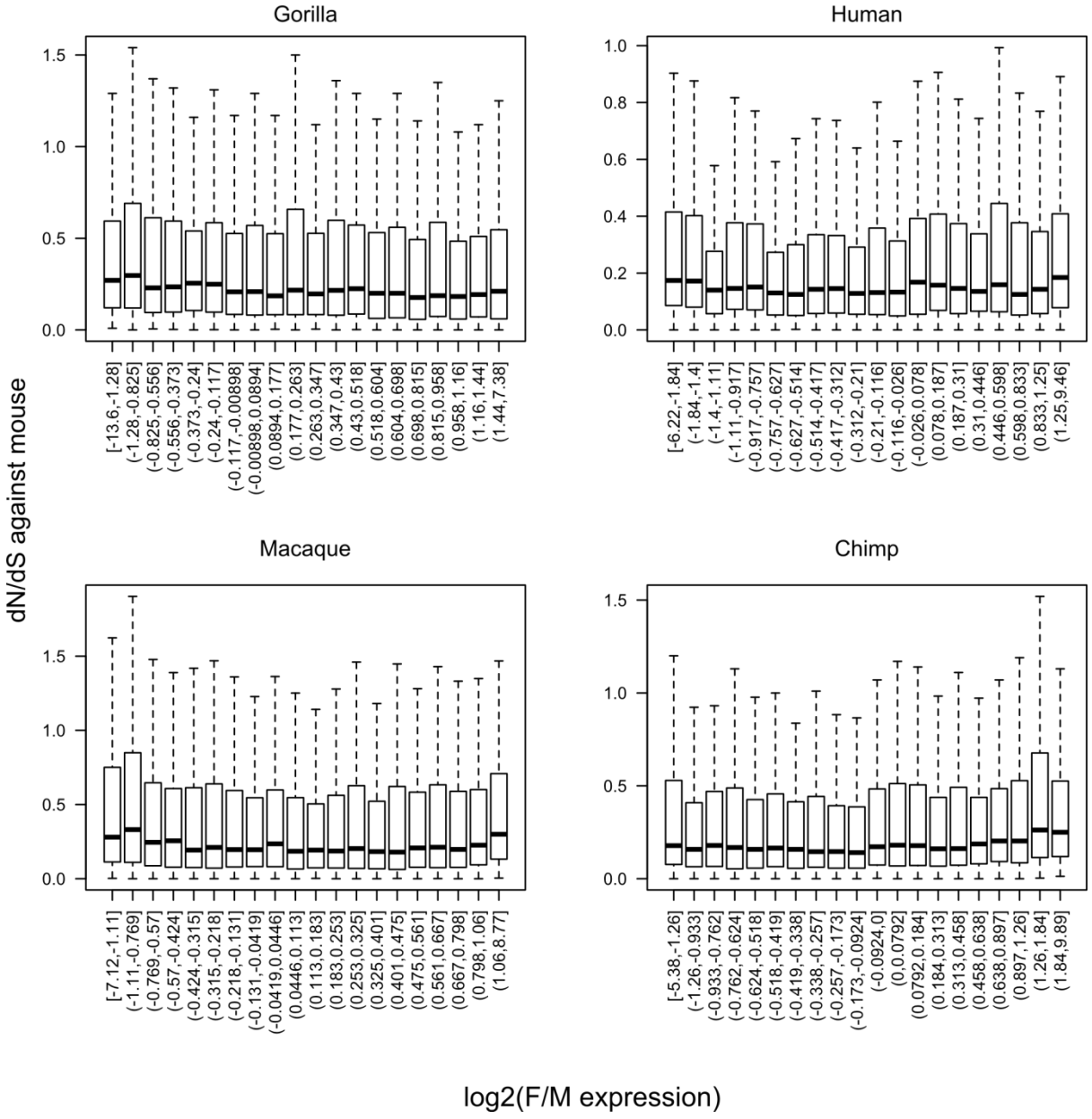


Figure S2-16. Boxplot showing trend between the sex bias of the expression of each gene in cerebellum (X-axis) and its dN/dS computed comparing against mouse (Y-axis) for 4 primate species with complete genome assembly. Boxes denote interquartile ranges, lines denote medians, and whiskers denote 1.5 times the interquartile range. Each box in the boxplot was constructed by dividing the genes into 20 quantiles according to the strength of the sex biased in their gene expression.

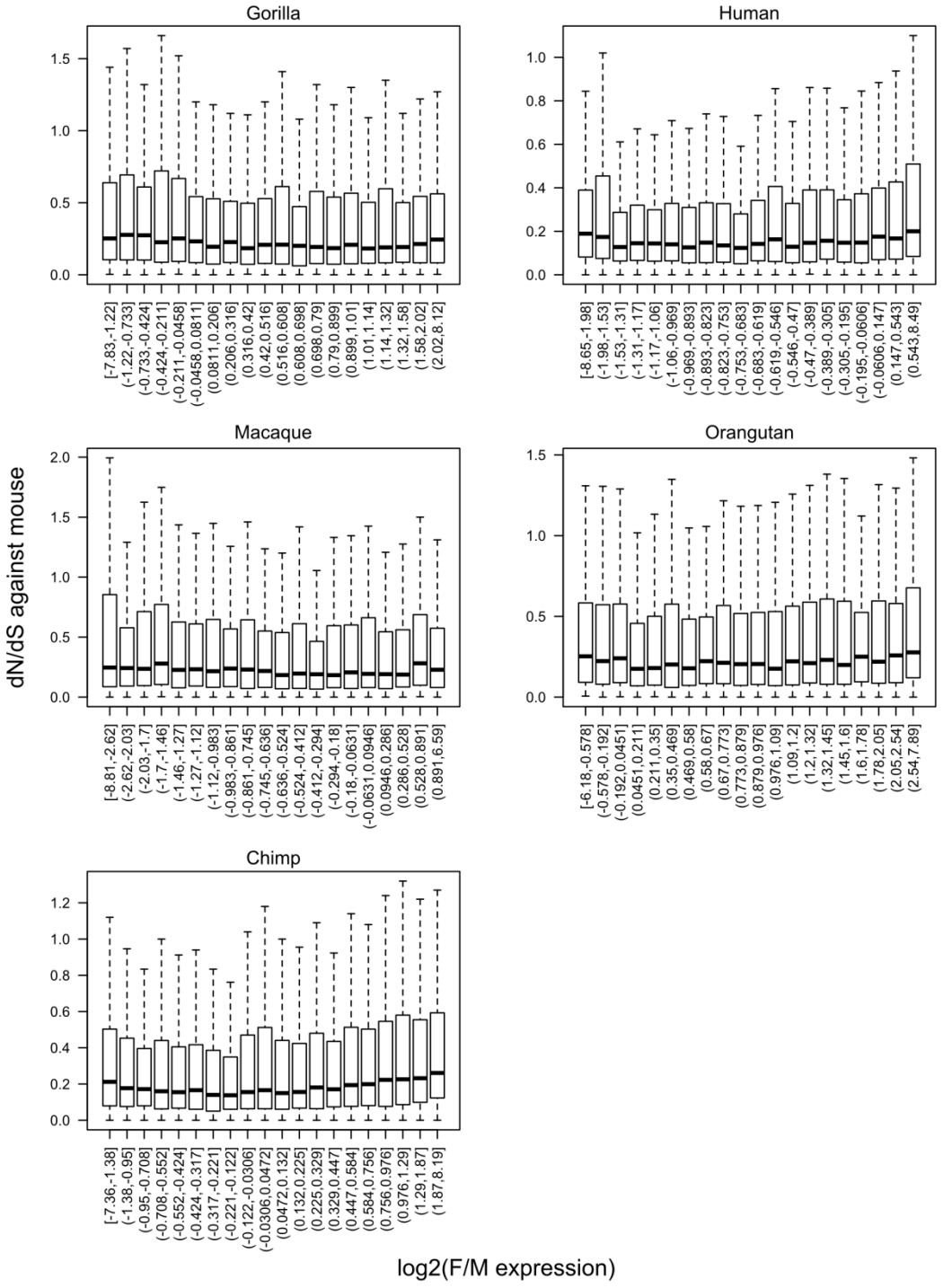


Figure S2-17. Boxplot showing trend between the sex bias of the expression of each gene in heart (X-axis) and its dN/dS computed comparing against mouse (Y-axis) for 5 primate species with complete genome assembly. Boxes denote interquartile ranges, lines denote medians, and whiskers denote 1.5 times the interquartile range. Each box in the boxplot was constructed by dividing the genes into 20 quantiles according to the strength of the sex biased in their gene expression.

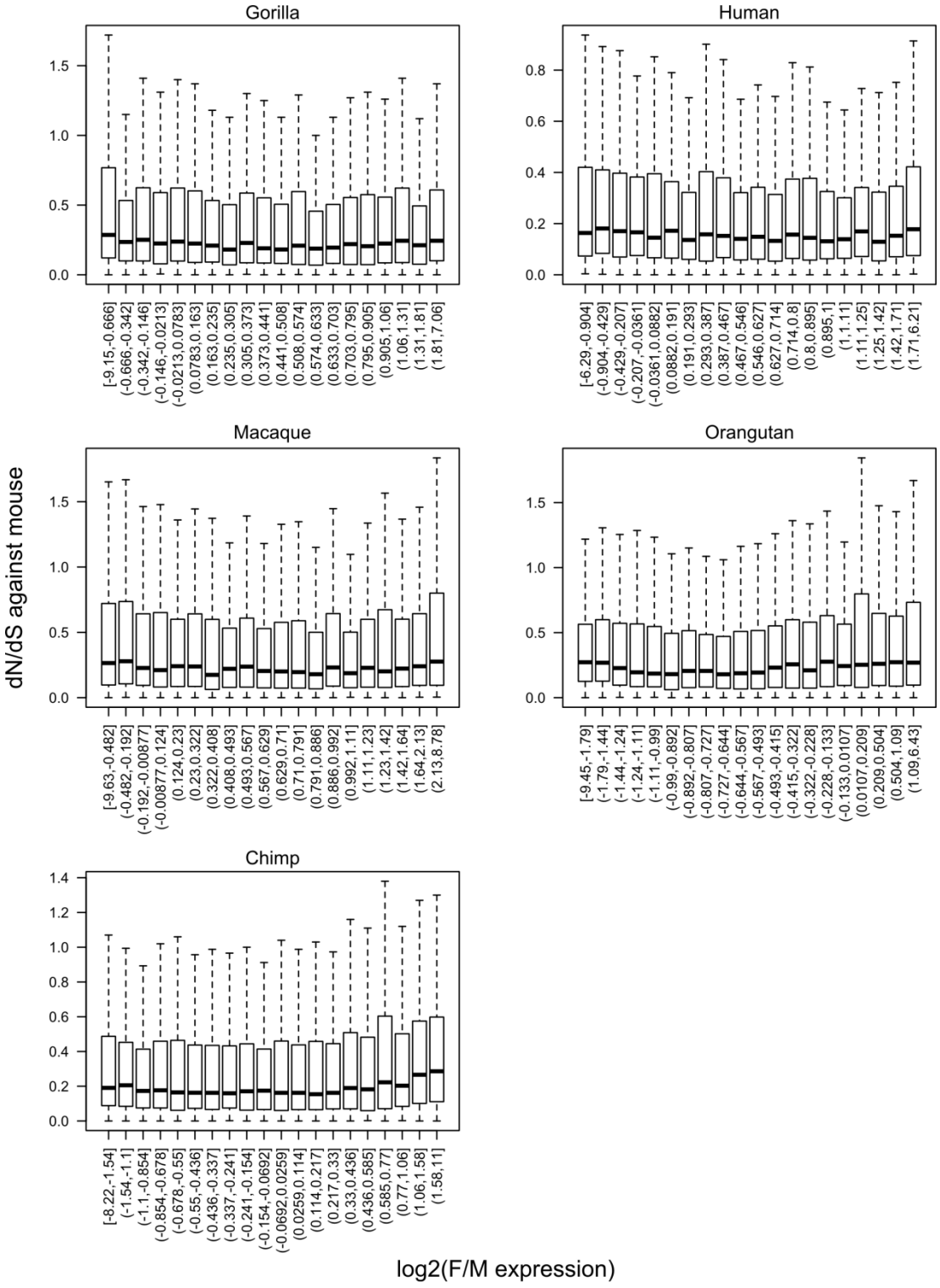


Figure S2-18. Boxplot showing trend between the sex bias of the expression of each gene in kidney (X-axis) and its dN/dS computed comparing against mouse (Y-axis) for 5 primate species with complete genome assembly. Boxes denote interquartile ranges, lines denote medians, and whiskers denote 1.5 times the interquartile range. Each box in the boxplot was constructed by dividing the genes into 20 quantiles according to the strength of the sex biased in their gene expression.

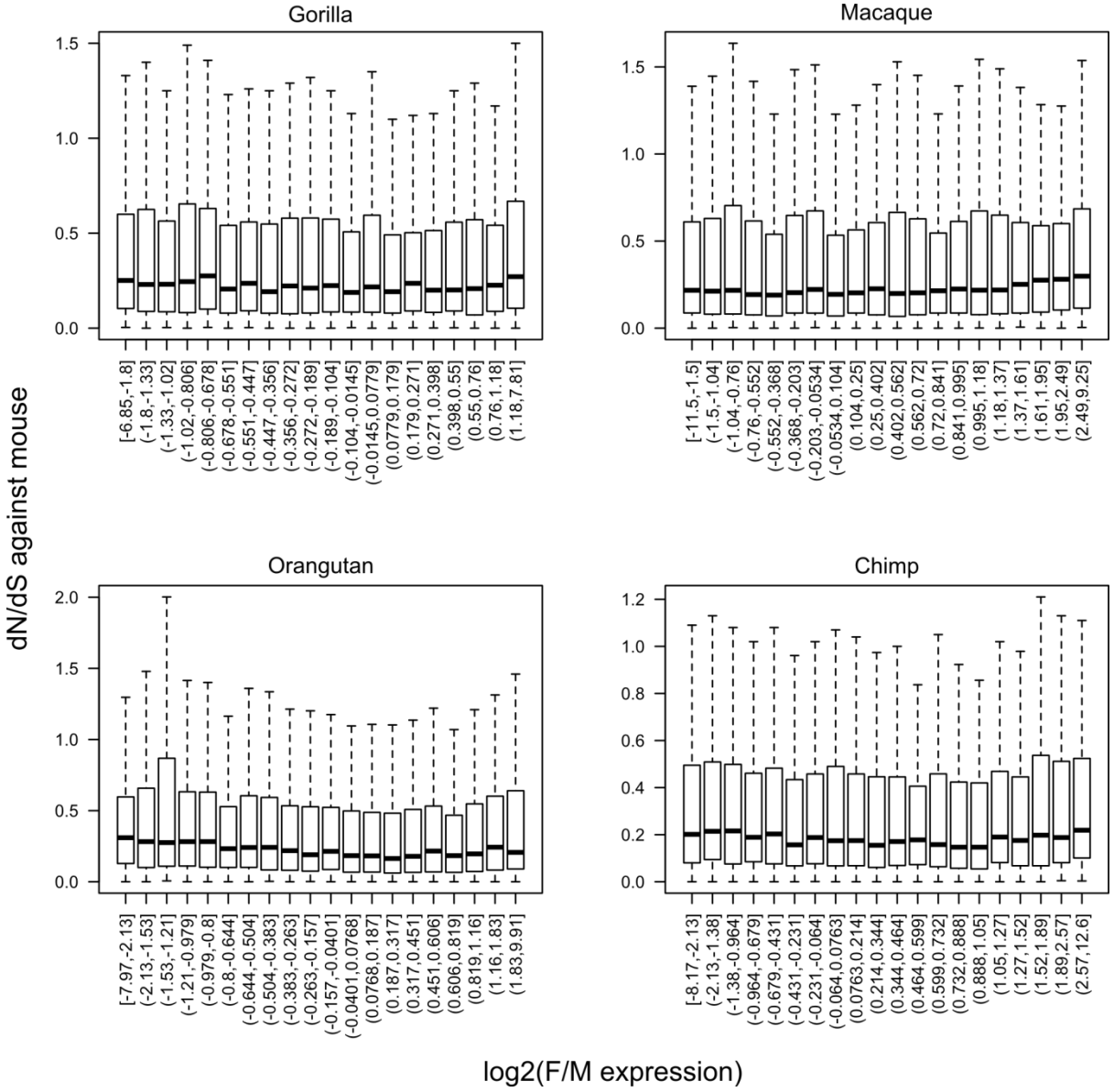


Figure S2-19. Boxplot showing trend between the sex bias of the expression of each gene in liver (X-axis) and its dN/dS computed comparing against mouse (Y-axis) for 4 primate species with complete genome assembly. Boxes denote interquartile ranges, lines denote medians, and whiskers denote 1.5 times the interquartile range. Each box in the boxplot was constructed by dividing the genes into 20 quantiles according to the strength of the sex biased in their gene expression.

Chapter 3 Conservation of testis over-expressed genes in *Drosophila*

3.1 Abstract

In eukaryotic genomes, genes are not randomly distributed. It has been discovered that, in *Drosophila melanogaster*, testis over-expressed genes tend to form clusters. Although the clustering is highly significant, a recent analysis has shown that there is little conservation of co-expressed genes in this genome, suggesting that clustering might result from neutral processes or transient selective pressures. Here, using testis-overexpressed gene clusters, we performed a comparative analysis to investigate whether testis clusters are conserved through the evolution of *Drosophila*. We show that testis over-expressed gene clusters are associated with higher degrees of gene loss and synteny breakage compared to genes outside the clusters, particularly in the most distant species analysed relative to *D. melanogaster*. Gene loss and synteny breaks are more common among gene pairs where both members are testis over-expressed in species most distant from *D. melanogaster*. The proportion of linkage breaks is higher as a result of chromosomal rearrangement. Nevertheless, clustered genes tend to reside in lower recombination rate regions. In addition, our results demonstrate that gene order among testis over-expressed genes evolves rapidly with linkage acquired later in evolution and higher linkage breaks for ancestrally linked testis genes.

3.2 Introduction

In bacterial genomes, about half of the genes are arranged in operon structures (Price et al. 2006). In contrast to earlier views, gene order in eukaryotes is now widely recognised to be non-random (Lee and Sonnhammer 2003; Hurst, Pal et al. 2004). Clustering of co-expressed genes on a genomic scale was first reported by Cho et al (1998) in the study of genes involved during the mitotic cell cycle of the budding yeast *Saccharomyces cerevisiae*. In mammals, gene order was then shown to be dependent on gene expression level or breadth of expression (Caron et al. 2001; Lercher et al. 2002; Trinklein, Aldred et al. 2004). Non-random expression of adjacent gene pairs has also been reported in plants (Cho et al. 1998; Williams and Bowles 2004; Chen, de Meaux et al. 2010). In *Drosophila melanogaster*, Spellman and Rubin (2002) found that over 20% of the genes tend to fall into groups which had members between 10 and 30 genes, with an average size of ~100kb.

The particular processes driving the non-random distribution of genes according to expression levels or breadth of expression appear to differ among species (Hurst, Pal et al. 2004). For example, in yeast, gene pairs which are transcribed divergently may be under selection (Seoighe, Federspiel et al. 2000). Genes controlled by the same sequence-specific transcription factor tend to be regularly spaced along the chromosome arms (Képès 2003). Highly clustered essential genes are in low recombination regions in yeast, and larger clusters have lower recombination rates suggesting that the maintenance of gene clusters might be under selection (Pal and Hurst 2003). This pattern does not seem to be universal, however, as in *Drosophila* there was no clear evidence that housekeeping clusters had low recombination rates (Weber and Hurst 2011).

Whether the observed clustering of genes according to patterns of gene expression is functional remains is unclear. Examining gene order in 12 genomes in the *Drosophila* lineage, Weber and Hurst (2011) assessed the conservation of three classes of expression related clusters in *D. melanogaster* (small clusters of highly co-expressed genes, large clusters of functionally unrelated housekeeping genes, and adjacent and similarly expressed clusters as defined by Spellman and Rubin (2002) and found no evidence that any of the three cluster classes are preserved as synteny blocks over time. These results seem to support that, at least in this lineage, the observed patterns of gene clustering may not be maintained by selection.

Interestingly, however, genes with testis expression in *Drosophila melanogaster* have been shown to be in clusters throughout the genome (Boutanaev et al. 2002). Whether this clustering of testis expressed genes is functional or not remains to be assessed. Here using testis over expressed gene clusters we performed a comparative analyses to examine whether gene pairs within testis over expressed gene clusters are more likely to remain paired across time than gene pairs outside these clusters. We also compared the recombination rate between testis and non-testis genes, and analysed the gene expression in different stages of spermatogenesis.

3.3 Material and Methods

3.3.1 Genome data and expression data

Twelve *Drosophila* genomes sequences were downloaded from FlyBase (ftp://ftp.flybase.net/genomes/). Four species are assembled into chromosomes: *D. melanogaster*, *D. simulans*, *D. yakuba*, *D. pseudoobscura*, and scaffolds for eight

other *Drosophila* species genomes are available. Data on gene homology relationships for each *Drosophila* species were also downloaded from FlyBase.

Gene expression data were obtained from FlyAtlas (http://flyatlas.org/). Gene expression values were averaged among probes matching the same gene. Probes matching multiple genes were removed from further analyses.

Information for testis overexpressed genes was also gathered from FlyBase (ftp://ftp.flybase.net/). *Drosophila* male biased gene expression data were gathered from Zhang et al. 2007 (2007).

3.3.2 Removal of duplicated genes

Duplicated genes are expected to have similar expression patterns, and could give rise to a trivial clustering effect of tandem duplicates. Thus, for each duplicated pair, one of the two duplicated genes was randomly picked and deleted from the data set.

3.3.3 Identification of clusters

First, testis over-expressed genes were defined as those where the testis expression was at least five times higher than the average expression in other tissues. Cluster of testis over-expressed genes were defined as follows: Expression data per tissue was analysed and those genes with >5-fold higher expression in testis compared to other tissues were considered testis over-expressed. For each analysis genes were ordered along the chromosome. Genes were then ordered along the chromosome with each one analysed separately. A cluster was then defined as any

group of at least 3 genes where at least two thirds were classed as testis over expressed including the boundary genes of the cluster.

3.3.4 Identification of gene neighbours

Genomic coordinates for *D. melanogaster* genes were extracted from FlyBase (ftp://ftp.flybase.net/genomes/). Gene position information was obtained by searching fasta files for all genes. All genes were then ordered according to their chromosomal positions. Genes were defined as neighbours if there were no intervening genes between them on either strand. Adjacent gene pairs were then identified and classified according to whether they are located inside or outside of testis over expressed gene clusters, whether one or two members are over expressed in testis and according to their transcription orientation [convergent (+-), divergent (-+), or tandem (++ or --)].

All gene pairs with overlapping genes or containing genes which overlapped with other gene(s) outside the pair were removed from further analyses.

3.3.5 Identification of gene pair linkage among *Drosophila* species

Genomic locations of orthologous genes in other *Drosophila* species were used to determine orthologous presence and linkage in those genomes. Pairs were considered 'conserved' if both genes had orthologous in the original strand-wise arrangement and had no other genes inserted between them. Reversed pairs also possessed orthologous but had different strand-wise arrangement compared to the original pair. Pairs were considered unlinked if orthologous genes for both genes were found but the genes were separated by other genes, or found on different

chromosomes or scaffolds. Those gene pairs in which one or both genes were missing orthologous in a given species were also identified.

3.3.6 Defining ancestrally present genes and linkage

Genes were considered ancestrally present/linked if they are present/linked both in *D. melanogaster* and in any of the three most diverged *Drosophila* species: *D. mojavensis*, *D. virilis* or *D. grimshawi*. Meanwhile, genes were considered ancestrally unpresent/unlinked if they are present/linked in *D. melanogaster* but not present/linked in any of the most highly diverged species: *D. mojavensis*, *D. virilis* or *D. grimshawi*.

3.3.7 Definition of ancestrally testis expressed genes

Using male biased expression data, genes were considered ancestrally male biased if they are male biased in *D. melanogaster*and either in *D. mojavensis* or *D. virilis*; genes were ancestrally not male biased if they are not male biased in any of *D. melanogaster*, *D. mojavensis* or *D. virilis*.

Then we assumed a gene is a conserved testis over expressed gene if it is testis over expressed in *D. melanogaster* and ancestrally male biased; a gene is a novel testis express gene if it is testis over expressed in *D. melanogaster* and ancestrally not male biased; a gene is a non-testis express gene if it is not testis over expressed in *D. melanogaster* and not male biased in any *Drosophila* species.

3.4 Results

3.4.1 Identification of gene pairs and orthologous relationships

In order to assess the degree of synteny conservation among testis over expressed gene clusters, we analysed the linkage conservation of adjacent gene pairs of *Drosophila melanogaster* in 11 other *Drosophila* genomes.

A total of 122 clusters were identified in *D. melanogaster* with cluster sizes ranging from 3 to 37 genes. Table 3-18 summarizes the general number of different gene pairs in *D. melanogaster*. Of a total of 14,245 adjacent gene pairs identified in *D. melanogaster*, 5,932 were removed from the analyses as one or two genes were found to be overlapping with each other or with other genes.

For each pair of adjacent genes in *D. melanogaster*, presence and position of orthologous genes was assessed in 11 other *Drosophila* species with varying degrees of relatedness to *D. melanogaster* (see methods). Presence of orthologous genes and linkage conservation was assessed based on orthology annotations obtained from FlyBase (http://flybase.org/). Depending on whether orthologous sequences for each gene could be found in each of the examined genomes and if so whether the genes are found adjacent to each other, each *Drosophila*'s gene pair was classified into one of three categories. A pair of adjacent genes in *D. melanogaster* was considered to be conserved in other *Drosophila* genomes when both the individual sequences of genes and the strand-wise arrangement of the pair were conserved. If a pair's orthologous were found to be still adjacent but with a different strand-wise arrangement, then the pair was designated "inverted". Orthologous falling on different

chromosomes/scaffolds or separated by other intervening genes were considered "unlinked".

3.4.2 Comparison of gene neighbours conservation of gene pairs between *D. melanogaster* and other *Drosophila* species

We first proceeded to analyse the gene fate for 8313 non-overlapping adjacent gene pairs from *D. melanogaster* in 11 other *Drosophila* species. Figure 3-**20** shows orthologous presence of *D. melanogaster* gene's orthologous in other *Drosophila* species and synteny conservation in 11 *Drosophila* species. There is an increasing trend of linkage breakage and orthologous gene loss according to the divergence from *D. melanogaster*.

Next, we asked whether gene pair linkage conservation within testis over-expressed gene clusters is conserved compared to gene pairs outside of these clusters.

In order to test whether there are differences in linkage conservation for *D. melanogaster* adjacent gene pairs located within and outside testis clusters, we calculated the proportion of gene pair linkage loss and maintenance for both in-cluster and out-cluster gene pairs for which orthologous genes were identified in the 11 *Drosophila* species. Throughout all the 11 *Drosophila* species analysed (Figure 3-**21**A), there was a higher proportion of linkage conservation of gene pairs outside of testis over expressed gene clusters compared to inside clusters.

When examining linkage conservation dependent on expression status in testis tissue over-express genes, we found a similar pattern to that observed in clusters, with gene pairs is more likely to become apart for two testis over expressed

gene pairs in *Drosophila* species compared to pairs with no testis over expression (Figure 3-**21**B).

These findings suggest that linkage for in cluster / testis over express gene pairs tend to be broken compared without cluster / non testis over express gene pairs.Gene inversions in the transcription direction of one gene relative to its gene pair were also found to be higher inside clusters than outside.

3.4.3 Reconstruction of evolutionary events leading to the linkage breakage

Although the linkage loss is higher in testis clusters than outside clusters, it is not clear whether this is due to linkage breakage or the age at which the gene pairs formed during evolution. Therefore we assessed the linkage of gene pairs of *Drosophila* ancestor genome and reconstructed the evolutionary path of gene pair linkage acquisition. Genes were considered ancestrally unlinked if they are linked in *D. melanogaster* but not linked in any of the most highly diverged species: *D. mojavensis*, *D. virilis* or *D. grimshawi*. Figure 3-**23** shows the cumulative percentage of first detection of linkage for ancestrally unlinked gene pairs. The proportion of first linkage is always higher for both out cluster gene pairs and two over expressed gene pairs comparing to in cluster and non over expressed gene pairs, which suggests that the linkage among testis genes is acquired later in *Drosophila* evolution.

Then we asked whether for those ancestrally linked pairs in the *Drosophila* ancestral genome the likelihood of linkage breaks for in cluster pairs is any different than for out of cluster gene pairs. Genes were considered ancestrally linked if they are linked both in *D.melanogaster* and in any of the three most diverged

*Drosophila*species: *D.mojavensis*, *D. virilis* or *D. grimshawi*. We then examined the fate of ancestrally linked gene pairs in *D. pseudoobscura* (Figure 3-**24**) whose genome has been assembled into full chromosomes. We found that ancestrally linked genes within testis over expressed gene clusters are more likely to lose their linkage compared to those pairs outside the testis gene clusters in *D. pseudoobscura*.

To identify the likely mechanism by which gene linkage was lost, we assessed the distance between *D. melanogaster* gene pairs which were ancestrally linked in *D. pseudooscura*. We found that, for those gene pairs for which linkage was lost, there is an excess of chromosomal rearrangements (Figure 3-**25**).

For those gene pairs which are ancestrally linked but changed their partners in *D. pseudoobscura*, we analysed their new neighbours and found that there are no significant preferences for either non-over expressed gene or over expressed gene linking to an over expressed gene or non-over expressed gene (Table 3-19). For the over expressed genes, 3 cases out of 32 (9%) changed their neighbours from an out cluster gene into an in cluster gene, which is consistent with the random expectation of 13 per cent. And, for over expressed genes linked to an in cluster gene, their replaced new gene neighbours are always from the previous clusters rather than from other clusters.

3.4.4 Orthologous gene conservation comparison

Similar pattern was found for the orthologous gene presence throughout *Drosophila* species evolution. Testis gene orthologous presence is higher in closely related species and lower in distant related species (Figure 3-**26**).

We then examined the orthologous gene acquisition of ancestrally not present genes, and found that testis over expressed orthologous genes appeared earlier than non-testis over expressed genes throughout evolution (Figure 3-**27**).

3.4.5 Recombination rate

The essential gene clusters in yeast have been observed to reside in low recombination regions (Pal and Hurst 2003). Moreover, the analysis of co-expression clusters in *Drosophila* also demonstrated that recombination is associated with gene order.

Therefore, recombination rate was considered for the gene pair linkage loss between testis over expressed genes and non-testis over expressed genes. Recombination rate in regions of testis genes is lower than in regions of non-testis genes (testis mean = 1.29, non-testis mean = 1.39, P=0.06). We also examined whether in cluster genes are in regions of higher recombination. The comparison shows that in cluster genes have a lower recombination rate than out cluster genes (in cluster mean = 1.17, out cluster mean = 1.37, $P < 0.001$).

Then we asked whether genes which change neighbours are located in regions of higher recombination. By analysing recombination rate in three *Drosophila* species (*D.simulans*, *D.sechellia* and *D.pseudoobscura*), we found no significant differences between linked genes and unlinked genes (Table 3-20).

3.4.6 Spermatogenesis stages

Next we ask whether genes associated with different stages of spermatogenesis are associated with different chances of losing their linkage. Genes

with different stages of spermatogenesis were analysed, and the result indicates that there is a higher linkage loss in the stages of mitotic and mitotic-meiotic throughout the evolutionary distance (Figure 3-**28**).

3.4.7 Does a gene acquire its testis function by moving next to a conserved testis gene?

The results show that testis over expressed genes in *D. melanogaster* appeared early through evolution, but the linkage was acquired late. Whether or not an ancestrally non testis over express gene acquired a testis over expression due to its moving to a testis over expressed gene is unclear. Thanks to the published *Drosophila* male biased expression data, we can use seven *Drosophila* species to address this question.

We defined ancestrally male biased gene as gene which is male biased in *D. melanogaster* but not in *D. mojavensis* or *D. virilis*; ancestrally not male biased gene as gene which is not male biased in any of *D. melanogaster*, *D. mojavensis* or *D. virilis*. Then we assumed a gene is a conserved testis express gene if it is testis over expressed in *D. melanogaster* and ancestrally male biased; a gene is a novel testis express gene if it is testis over expressed in *D. melanogaster* and ancestrally not male biased; a gene is a non-testis express gene if it is not testis over expressed in *D. melanogaster* and not male biased in any *Drosophila* species.

Again, we traced the conserved/novel testis gene pair acquisition in 11 *Drosophila* species (Figure 3-**29**). Linkage formation for novel testis gene pairs appears late comparing with conserved testis gene pairs. When assessing the linkage break for ancestrally linked pairs in *D. pseudoobscura*, we found that linkage break

is higher for novel testis gene pairs (Figure 3-**30**). However, the sample size is very small and it is not statistically significant.

Therefore, next we ask whether a novel testis gene acquired its testis over express by moving next to a conserved testis gene. The distance for male biased gene in *D. yakuba* is significantly shorter than that in *D.pseudoobscura*. However, the distance for not male biased gene in *D. yakuba* is also significantly shorter than that in *D. pseudoobscura*. There are totally 89 novel testis genes which is adjacent to a conserved testis gene in *D. melanogaster*. For these genes, there is no significant difference between male biased gene and not male biased gene in *D. yakuba*. Therefore, we made a 1000 times randomisation and calculated the ratio of average distance in *D. yakuba* over average distance in *D. pseudoobscura* for not male biased genes. A one-sample t-test was conducted to compare distance for male and not male biased gene to its nearest conserved testis gene in *D. pseudoobscura* and *D. yakuba*. There was a significant difference in the ratio for *D. pseudoobscura* and *D. yakuba*; $t(99) = 5.869$, $p < 0.001$. The result suggests that novel testis gene moves closer to a conserved testis gene in *D. yakuba* than not male biased genes.

3.5 Discussion

In *Drosophila*, genes with testis over expression are non-randomly distributed and observed to be clustered (Boutanaev, et al. 2002). Here, we examined the conservation of gene order for testis gene pairs using 11 *Drosophila* species.

We found that the linked gene pairs in *Drosophila melanogaster* often broke their linkage in other *Drosophila* species. In yeast, the transcriptional orientations of a gene pair are often different between relative species (Seoighe, et al. 2000). In

consistent with the previous study, testis gene pairs in *Drosophila* also exhibit transcription direction changes.

Previous study in yeast showed that highly clustered essential genes are in low recombination regions (Pal and Hurst 2003). However, in the study of *Drosophila* housekeeping genes, Weber and Hurst (2011) found no clear evidence that housekeeping clusters had low recombination rates. In our study, we also found a significantly low recombination rate in regions of testis genes. Nevertheless, the in cluster genes are in low recombination rate region compared with out cluster genes. However, when we examined whether gene which changes its neighbour is located in regions of higher recombination rate, we didn't find any relations between gene pair linkage break and recombination rate.

In yeast, Seoighe et al. (2000) proposed that small scale inversion plays an important role for adjacent gene linkage broken. Our result also showed that in addition to inversion, inter-chromosome rearrangement could also contribute to high linkage break of gene clusters.

The comparison of *Drosophila melanogaster* to other species indicates that testis clusters in *D. melanogaster* are not conserved in others. However, is it the case for ancestrally linked gene pairs? In order to assess the fate of ancestral gene pairs, we inferred the expression status for ancestral genes with three most distant species in *Drosophila* lineage, *D. mojavensis, D. virilis*and *D. grimshawi*. The observation shows that testis genes appeared earlier, but linkage among them was acquired later in evolution. This can be explained by a non-testis gene either gained its testis function first and then moved next to another testis gene, or moved next to a testis gene first then gained its testis function. We explored this question by using male

biased gene expression data. The findings suggest that linkage for conserved testis gene pairs is higher through evolution. The findings also suggest that there is evidence that novel testis genes are moving closer to conserved testis genes than non-testis genes.

3.6 References

Boutanaev, A. M., A. I. Kalmykova, et al. (2002). "Large clusters of co-expressed genes in the Drosophila genome." Nature**420**(6916): 666-669.

Caron, H., B. v. Schaik, et al. (2001). "The Human Transcriptome Map: Clustering of Highly Expressed Genes in Chromosomal Domains." Science**291**(5507): 1289-1292.

Chen, W.-H., J. de Meaux, et al. (2010). "Co-expression of neighbouring genes in Arabidopsis: separating chromatin effects from direct interactions." BMC Genomics**11**(1): 178.

Cho, R. J., M. J. Campbell, et al. (1998). "A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle." Molecular Cell**2**(1): 65-73.

Hurst, L. D., C. Pal, et al. (2004). "The evolutionary dynamics of eukaryotic gene order." Nat Rev Genet**5**(4): 299-310.

Képès, F. (2003). "Periodic Epi-organization of the Yeast Genome Revealed by the Distribution of Promoter Sites." Journal of Molecular Biology**329**(5): 859-865.

Larracuente, A. M., T. B. Sackton, et al. (2008). "Evolution of protein-coding genes in Drosophila." Trends in Genetics**24**(3): 114-123.

Lee, J. M. and E. L. L. Sonnhammer (2003). "Genomic Gene Clustering Analysis of Pathways in Eukaryotes." Genome Research**13**(5): 875-882.

Lercher, M. J., A. O. Urrutia, et al. (2002). "Clustering of housekeeping genes provides a unified model of gene order in the human genome." Nat Genet**31**(2): 180-183.

Pal, C. and L. D. Hurst (2003). "Evidence for co-evolution of gene order and recombination rate." Nat Genet**33**(3): 392-395.

Price, M. N., A. P. Arkin, et al. (2006). "The Life-Cycle of Operons." PLoS Genet**2**(6): e96.

Seoighe, C., N. Federspiel, et al. (2000). "Prevalence of small inversions in yeast gene order evolution." Proceedings of the National Academy of Sciences**97**(26): 14433-14437.

Spellman, P. and G. Rubin (2002). "Evidence for large domains of similarly expressed genes in the Drosophila genome." Journal of Biology**1**(1): 5.

Trinklein, N. D., S. F. Aldred, et al. (2004). "An Abundance of Bidirectional Promoters in the Human Genome." Genome Research**14**(1): 62-66.

Weber, C. and L. Hurst (2011). "Support for multiple classes of local expression clusters in Drosophila melanogaster, but no evidence for gene order conservation." Genome Biology**12**(3): R23.

Williams, E. J. B. and D. J. Bowles (2004). "Coexpression of Neighboring Genes in the Genome of Arabidopsis thaliana." Genome Research**14**(6): 1060-1067.

Zhang, Y., D. Sturgill, et al. (2007). "Constraint and turnover in sex-biased gene expression in the genus Drosophila." Nature**450**(7167): 233-237.

3.7 Tables and Figures

Table 3-18 General information of gene pair in *D. melanogaster*

D. melanogaster	**No. of gene pairs**
Genes overlapped pairs	2,266
Gene involved in overlap pairs	3,666
Usable gene pairs	8,313
Total	14,245

Table 3-19. New neighbours for ancestrally linked gene pairs in D. pseudoobscura

New neighbour	**N[1]**	**O[2]**	**expected**
O	10.2%	16.7%	13.0%
N	89.8%	83.3%	87.0%

[1]Non over-expressed gene
[2] Over-expressed gene

Table 3-20 Recombination rate of linked/unlinked gene pairs between species

	Linked	**Unlinked**	**p-value**
Dmel_Dsim	1.383928	1.34087	0.181
Dmel_Dsec	1.401752	1.322685	0.013
Dmel_Dpse	1.378492	1.412322	0.305

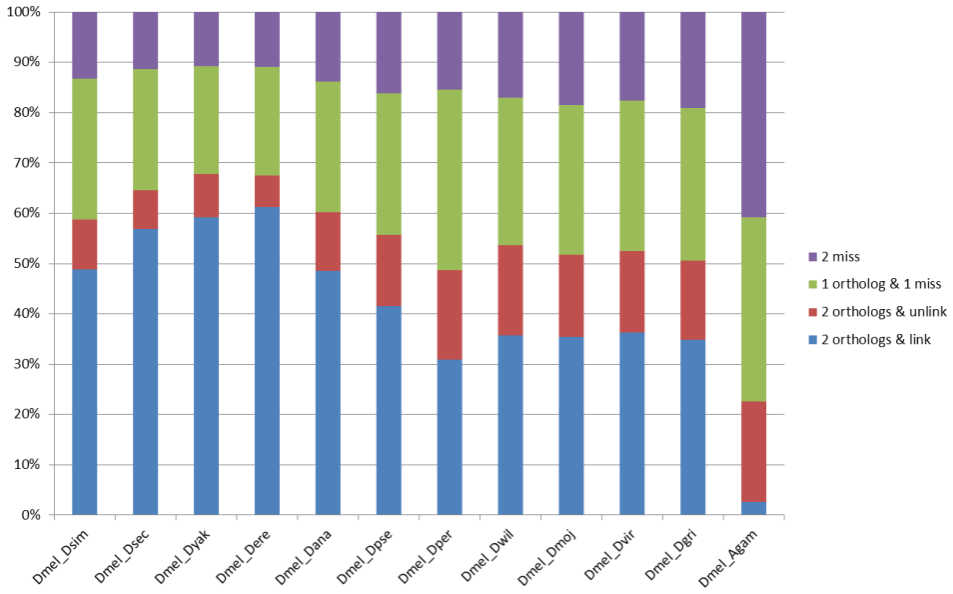
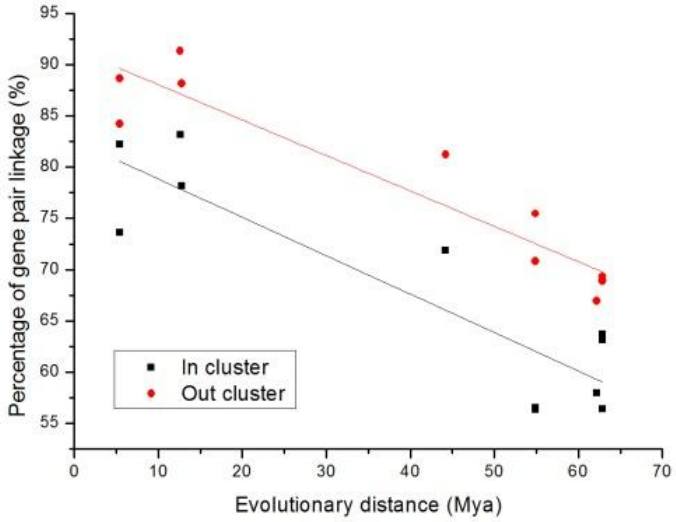
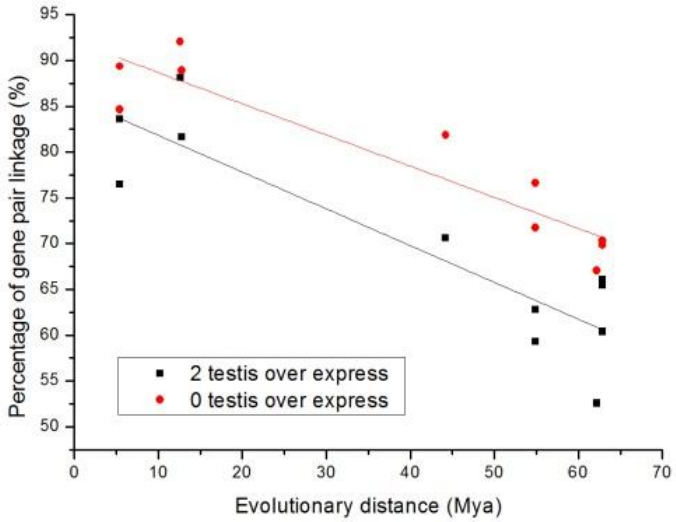


Figure 3-20 Orthologous gene presence for *Drosophila melanogaster* adjacent gene pairs in other *Drosophila* species. Bars represent percentage of gene pairs with two orthologous present and linked (blue), two orthologous present and linked (red), one orthologous present (green) and, no orthologous present (purple).

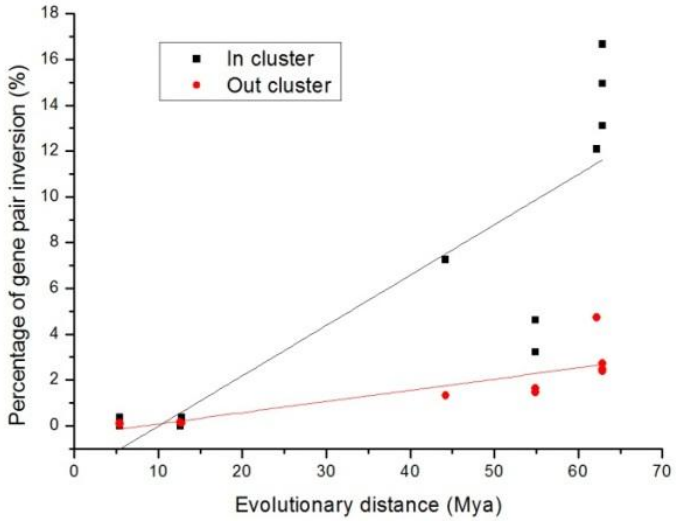


A

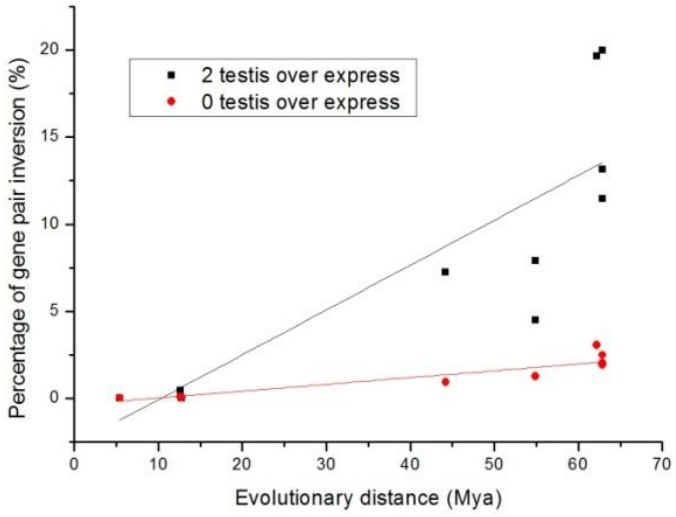


B

Figure 3-21 Gene pair linkage conservation in *Drosophila* genus. A) Gene pair linkage of in/out testis clusters. B) Gene pair linkage of testis over expressed gene pairs. Horizontal axis represents the evolutionary distance (Mya) for each species to *D. melanogaster*. Dots represent the percentage of gene pair linkage between *D. melanogaster* and other *Drosophila* species.

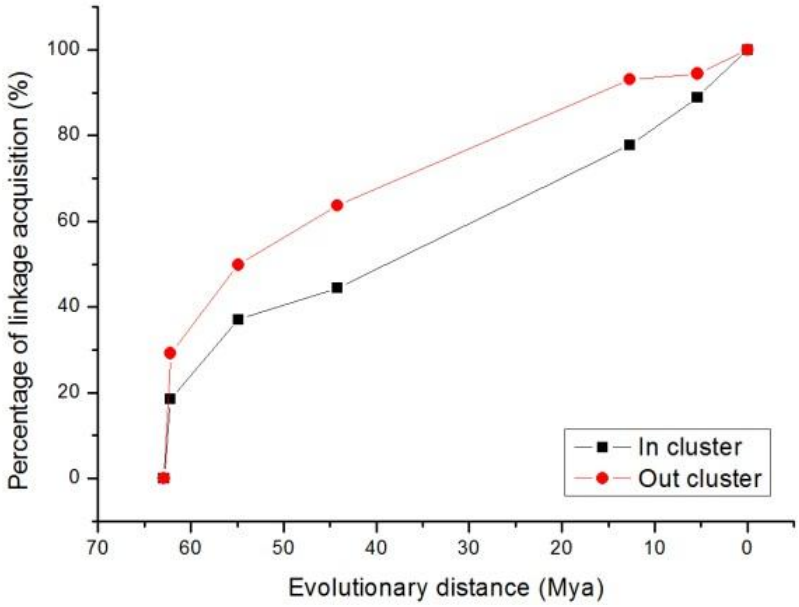


A

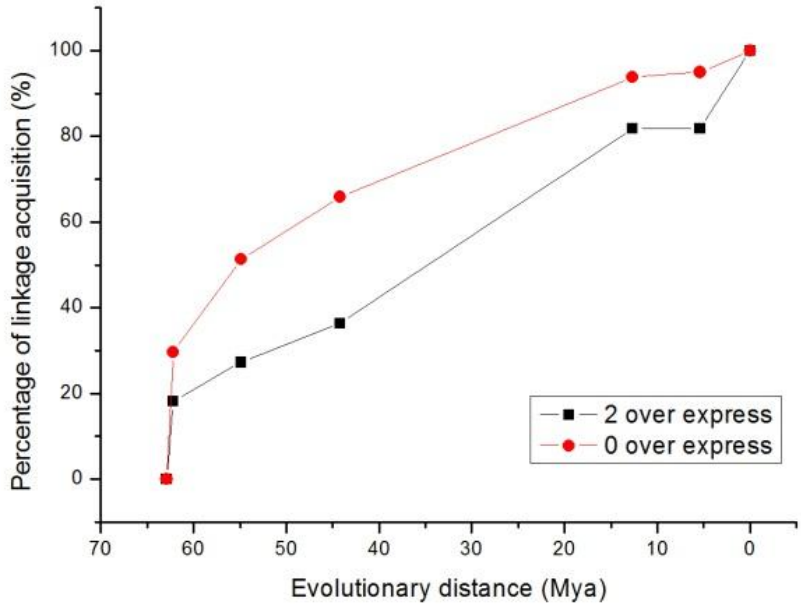


B

Figure 3-22 Percentage of gene pair inversion in each *Drosophila* species. A) gene inversion for in / out cluster gene pairs; B) gene inversion for testis over expression. Horizontal axis represents the evolutionary distance (Mya) for each species to *D. melanogaster*. Dots represent the percentage of gene pair linkage between *D. melanogaster* and other *Drosophila* species.



A



B

Figure 3-23 Accumulative percentage of linkage gain in *Drosophila* genus from *Drosophila* ancestor genome. A) Gene pair linkage acquisition in/out clusters; B) Linkage acquisition for testis expressed genes. Horizontal axis represents the evolutionary distance (Mya) for species to *D. melanogaster*. Dots represent the percentage of gene pair linkage between *D. melanogaster* and other *Drosophila* species.

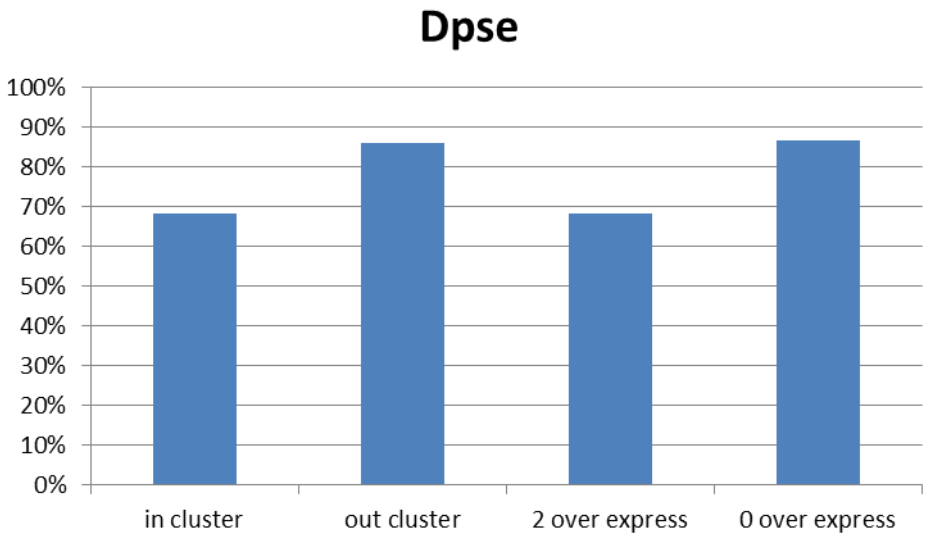
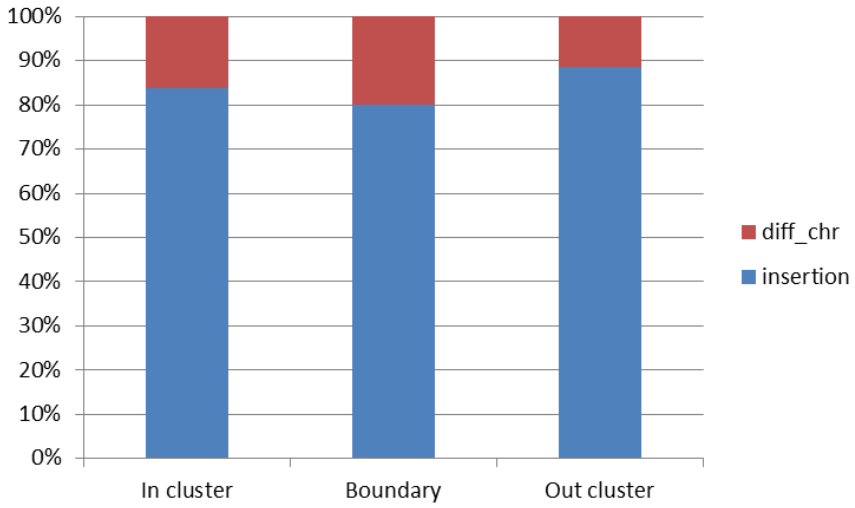
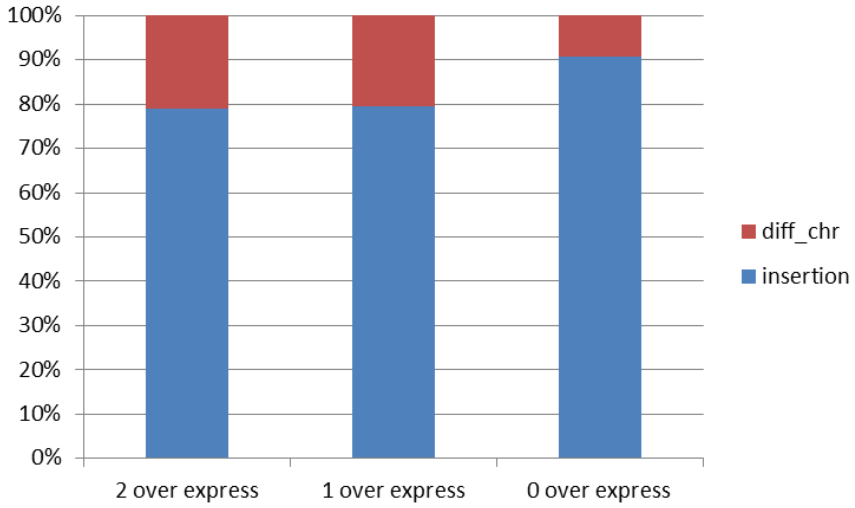


Figure 3-24 Linkage conservation of ancestor gene pairs in *D. pseudoobscura*. Each bar represents the percentage of gene pairs still linked in *D. pseudoobscura* compared to the *Drosophila* ancestor gene pairs.

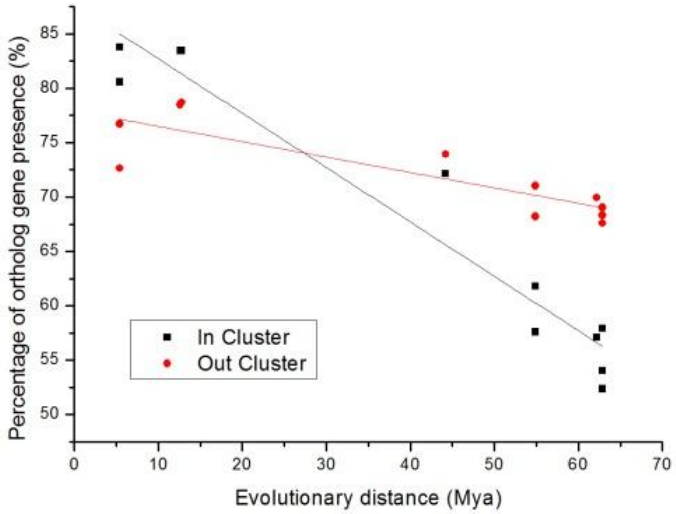


A

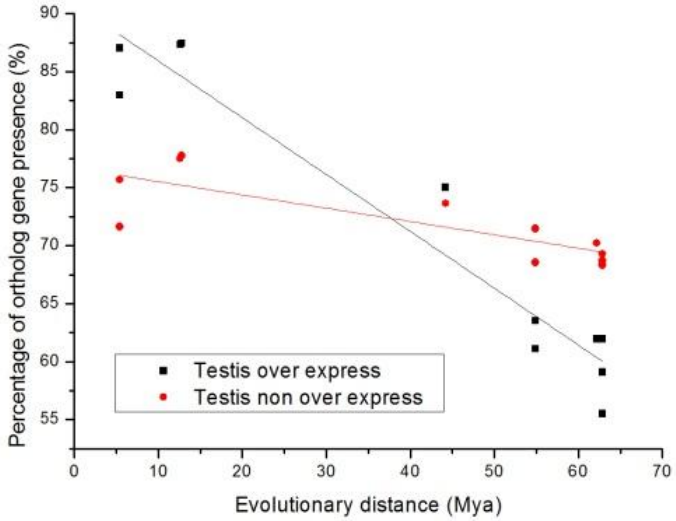


B

Figure 3-25 Gene insertion for ancestor linked gene pairs in *D. pseudoobscura*. A) Gene insertion for gene pairs in/out testis clusters; B) Gene insertion for gene pairs with different testis expression.

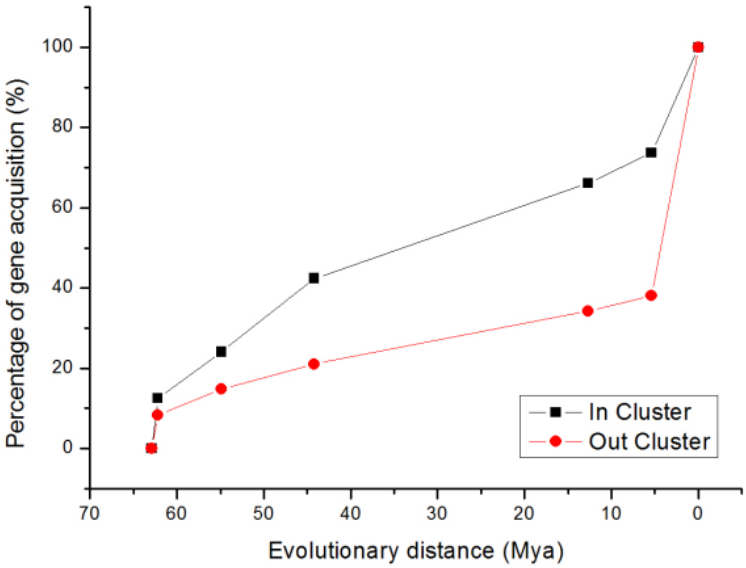


A

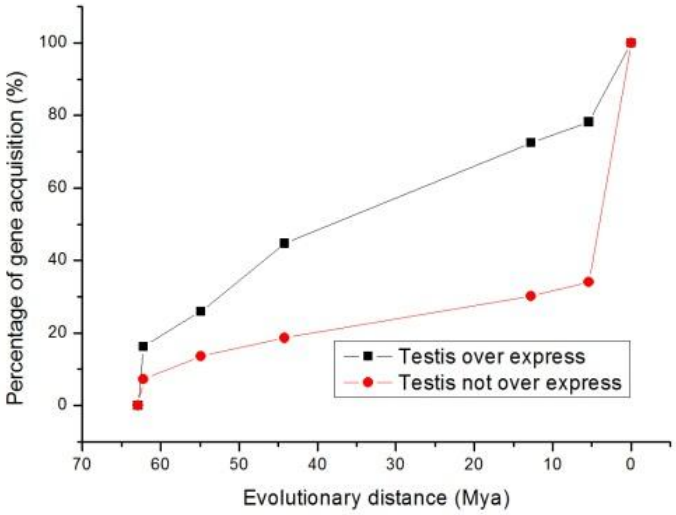


B

Figure 3-26 Orthologous gene presence of *D. melanogaster* in other *Drosophila* species. A) Orthologous gene presence in/out testis clusters; B) Gene presence of testis over expressed gene. Horizontal axis represents the evolutionary distance (Mya) for each species to *D. melanogaster*. Dots represent the percentage of orthologous gene presence in each *Drosophila* species relative to *D. melanogaster*.



A



B

Figure 3-27 Accumulative percentage of orthologous gene gain in *Drosophila* genus from *Drosophila* ancestor genome. A) Orthologous gene obtained in/out testis clusters; B) Orthologous gene obtained for different testis expression. Horizontal axis represents the evolutionary distance (Mya) for species to *D. melanogaster*.

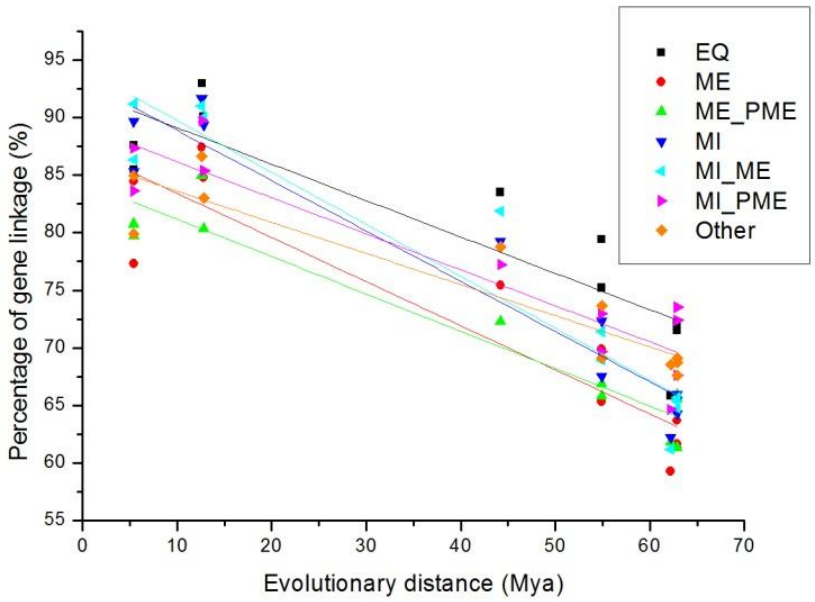


Figure 3-28 Linkage loss in different stages of spermatogenesis.

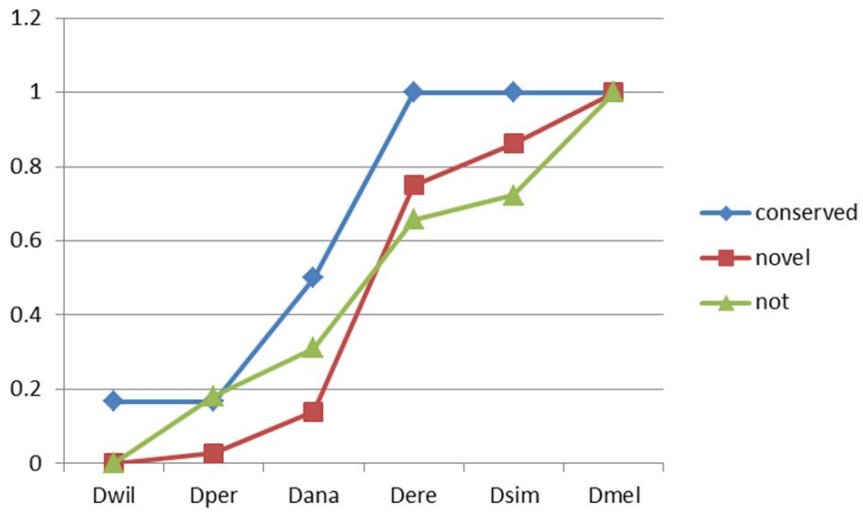


Figure 3-29. Linkage acquisition for conserved/novel testis gene pair. Y-axis represents the accumulative percentage of linked gene pair.

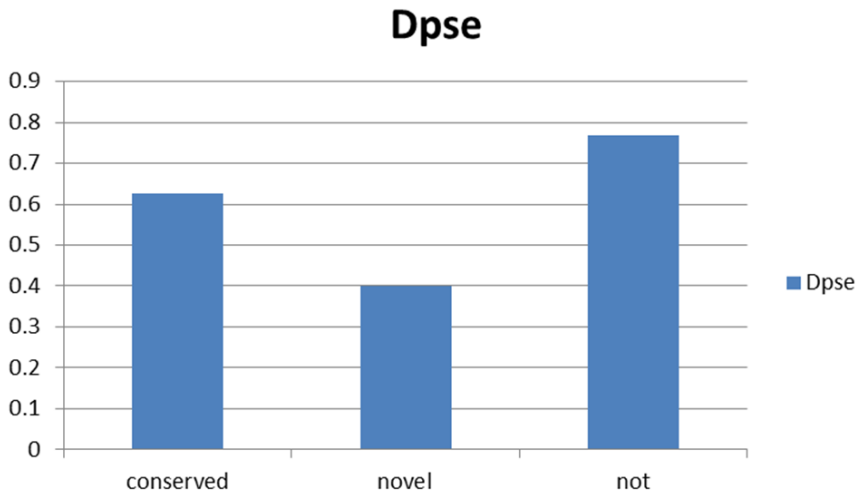


Figure 3-30. Proportion of linkage for conserved/novel testis gene pair in *D. pseduoobscura*.

Chapter 4 Genetic divergence and parallel responses to selection for early flowering in *Arabidopsis thaliana*

4.1 Abstract

Understanding how population allele frequencies change in response to selective pressures can provide valuable insights into the molecular basis of adaptation. Most studies aimed at uncovering the genomic basis of adaptation rely on either sequence comparison across different species or apply quantitative trait locus (QTL) and genome wide association (GWA). While these types of studies have increased our understanding of adaptation, by studying populations in the wild, some of the genes identified might be mediating the evolution of other phenotypes which might co-vary with the phenotype of interest. Experimental selection set ups allow us to study changes in allele frequencies in response to selective constraints in a controlled manner. Here we analyse changes in allele frequencies in the model plant species *Arabidopsis thaliana* in response to selection for early flowering under two dark/light cycles resembling those found in winter and spring conditions. We find that there are significant changes in allele frequencies after only six generations both when analysing individual SNPs and examining global genome patterns of changes in allele frequencies. We further find that although there are few SNPs with consistently significant changes across all selected lines, there are significant parallel changes in SNP frequencies on a genomic scale when analysing winter and spring conditions but not when analysing the similarities transcending both growing

conditions. Together these results suggest that using experimental selection setups can allow us to identify SNPs associated with adaptation in a specific trait and we show that there is evidence for widespread changes in allele frequencies in *A. thaliana* in response to selection for early flowering.

4.2 Introduction

Identifying the genetic changes that mediate adaptation is a major goal of evolutionary research. Despite the increasing accumulation of genomic data, how genomes adapt in response to a changing environment remains poorly understood. In plants, many phenotypic adaptations of relevance to understanding the effects of climate change or improving crop production are complex and involve changes at multiple loci (e.g., multiple loci determination). Using comparative analyses to compare genomes from diverse species can be used to identify genes which might have played an important role in the evolution of a particular trait (Li and de Magalhães, 2013, de Magalhães and Church, 2007). While this approach can yield important information about the genomic basis of long term evolution of a particular phenotype, and we expect it to become more widely used as more sequenced genomes become available, this approach can be biased by the fact that some of the genes identified might be supporting the evolution of other phenotypes which might be co-evolving with the trait of interest. In addition, the use of multispecies analyses might be less useful in identifying genes that mediate change and adaptation in a species specific manner.

The identification of loci driving phenotypic adaptation within a species often involves using quantitative trait locus (QTL) analysis or genome-wide association

(GWA) analyses (Bergelson and Roux, 2010, Brachi et al., 2010, Brown et al., 2004, Buckler et al., 2009, Kover et al., 2009a, Ross-Ibarra et al., 2007). However, these approaches may not provide a comprehensive interpretation of how selection results in phenotype variation by altering allele frequencies, as they only identify the loci which explain adaptive phenotype variation. The limitation originates from statistical issues, which constrain the quantity of loci identified by QTL analysis, and bias in identifying functional SNP by intermediate frequency in the GWA studies. Furthermore, the loci that explain most of the phenotype variation in a certain trait may not have response to selection, for the reason that they also confer deleterious pleiotropic effects, which can reduce the viability of the plant, or because the effects depend on the interaction with other loci (epistasis).

Thanks to the vast efforts made by molecular, developmental and evolutionary biologists, flowering time in *Arabidopsis thaliana* becomes an ideal trait connecting selection at phenotype level and response at molecular level. Expectation of strong selection is given to flowering time. Meanwhile, in light of the fact that the optimal reproduction time relies on the local environment in many species, flowering time is a key feature for local adaptation.

Flowering time has also been shown to impact other traits including physiological plasticity, resistance to herbivores, and inflorescence structure(Zhang and Lechowicz, 1994, Weinig et al., 2003, Scarcelli et al., 2007). In the meantime, huge variation of flowering time (from 12 days to six months) among natural *Arabidopsis* accessions has been observed under many different environmental conditions (Brachi et al., 2010, Atwell et al., 2010, Wilczek et al., 2009).

Although a continuous variation has been observed in flowering time, two extreme strategies can be identified: 1) “rapid-cycling strategies” (sometimes also referred to a “summer-annuals strategy”), which germinate in the Spring, Summer, or even Fall and flower within the same growing season; and 2) “winter-annual strategies”, which usually germinate in the Fall, overwinter as a rosette, and flower in the next growing season. The success for summer annuals depends on whether they are fast enough to complete their process of seed producing. In contrast, the success for winter annuals depends on whether they can precisely sense the end of winter conditions.

Flowering time is also an ideal trait for studying the molecular mechanism of selection response, due to the fact that the genetic basis of flowering time has been extensively investigated. More than 100 genes have been discovered in *Arabidopsis thaliana* by the analysis of artificial mutants. These genes are distributed in four interacting pathways: the “photoperiod pathway”, which mediates the early flowering response to longer days; the “vernalization pathway”, which mediates the response to cold exposure; the “Gibberelin pathway”, which mediates hormonal signalling; and the “autonomous pathway”. A lot of QTL analyses have been performed to study the genetic basis of natural variation in flowering time. More than 60 natural mutations, which contribute to flowering time, have been identified, (Alonso-Blanco et al., 1998, Ungerer et al., 2003, Brachi et al., 2010, Kover et al., 2009a). These studies highlight the complexity of the regulation of flowering time, among which different QTL has been identified under various genetic backgrounds, and the effects differ according to the conditions of plant growth.Therefore, identification of the SNP which respond to selection for flowering time in a complex

genetic background is helpful in the dissection of important traits of economic, and evolution.

A previous study has revealed that flowering time can evolve rapidly in an outbred population of *Arabidopsis thaliana*, reducing the time to flowering by ~2 standard deviations in six generations(Scarcelli and Kover, 2009). We have also shown that FRIGIDA (FRI)(Haerty et al., 2007), one of the genes with a large influence the flowering time, is involved in this response when selection is performed in the absence of chilling temperatures (i.e., under a simulated "spring-annual treatment"). However, FRI only explains a small part of flowering time change under the treatment; and FRI does not explain changes in flowering time under the simulation of winter-annual treatment.

Here we investigate which genetic factors mediated evolutionary changes in an experimental outbred population of *Arabidopsis thaliana*, by searching genome-wide for changes in allele frequencies in response to selection for early flowering. The extensive genomic and functional information available makes*A. thaliana* an excellent model organism to study the genetics of evolutionary adaptation, and currently the only higher plant species where such study has been carried out.

Here, by comparing allele frequencies of 1200 SNPs before and after selection under two different environments with three replicated populations, we investigate the genome-wide respond to selection in early flowering. Some previous studies have discussed genome-wide response to selection in fruit fly and domestic chicken (Johansson et al., 2010, Turner et al., 2011). Unlike these studies, we focus on short-time response, and whether the response is predictable and similar between replicates under same and different environment conditions. Understanding of early

stage response to selection and the environmental specificity to the response is crucial to apprehend the direct genetic consequence of environment changes.

4.3 Material and methods

4.3.1Allele frequency data for control and selected lines

Data for allele frequencies for a set of 1200 SNPs described in Kover et al (2009a) for lines under selection for early flowering and controlswere kindly shared by the Kover lab via personal communication. The experimental protocol used to obtain these data is similar to that described in Scarcellietet al (2007) and Kover et al (2009b). In brief, for the selection lines, only the 50 earliest flowering plants contributed to the next generations. Each of these 50 plants was randomly assigned to one cross as the male or a female parent, for a total of 25 crosses per line. In the control lines, 25 crosses per generation were performed, but the 50 parental plants were chosen randomly. Seeds from each cross were planted into eight pots, maintaining the line sizes at 200 plants every generation. This protocol was repeated for five generations. Crosses between lines with very different flowering times were accomplished with the help of pollen storage (for more details, see Scarcelli and Kover, 2009).

4.3.2Dendrogram construction

To construct the dendrogram tree, firstly a correlation matrix wascalculated between control and selected lines. Then a distance matrixwas calculated based on the correlation matrix between each winter/spring control and selected

lines(Supplementary Table S4-2 - S4-4). Finally the dendrogram were constructed based on thedistance matrix as implemented in the *pvclust*package of R.

4.3.3 Parallel divergence analysis

Overall divergence between lines: Using 10,000 bootstrap samples, sums of squared errors (SSE) were calculated using random samples of 20% of the SNPs between each line. To compare the distance of each control/selectedline diverging from parent populations within spring and winter, we counted the number of SSE in control(selected) line larger than selected(control) line and calculated the percentage. To compare the divergence between control lines and selected lines within spring and winter, we counted the number of SSE which is larger in control lines diverging from parent than in selected lines, and calculated the percentage. To compare the difference resulted by spring and winter conditions, we calculated the percentage which control/selected lines diverging from spring parent are larger than control/selected lines diverging from winter parent.

4.3.4 General data analysis

All statistical analyses were performed in the R software environment.

4.4 Results and discussion

4.4.1 Selection for early flowering significantly changed allele frequencies

Changes in allele frequency in control and selected lines for early flowering under winter and spring cycle day light pattern were analysed after six generations.

Changes in allele frequencies were examined for 1200 SNPs spread in the *A. thaliana* genome. The maximum Euclidean distance change in the allele frequency for any given SNP between control and selected lines was 0.47 for spring and 0.38 in winter conditions (Supplementary figure S4-34).

We then examined the global patterns of change in allele frequencies using a cluster analysis (Figure 4-31). We found that control and selected lines formed separate clusters (Monte Carlo simulation, $p = 0.0011$). Selected lines were further subdivided into distinct clusters depending on whether the plants were grown under spring or winter conditions (probability of the clustering of selected lines into a group distinct from controls and further subdivision into spring and winter conditions using Monte Carlo, $p = 0.000052$). Subdivision into winter and spring lines was not apparent among the control lines. Parent populations clustered with control lines suggesting that control lines diverged less in their allele frequencies compared to the selected lines. Given that seeds were randomly selected from a single seed pool to produce the two parent populations for spring and winter conditions, allele frequencies in these two are the most similar to each other.

In order to assess whether the patterns of clustering observed in Figure 4-1 is driven by the changes in those SNPs showing significant changes in allele frequencies or a more generalised pattern of change, clustering analysis was repeated using a reduced set of 65 SNPs.These SNPs showed a significant change in allele frequency in selected lines in at least two out of four tests after removing the gene with the lowest change in allele frequency from a total of eight pairs, which show a significant degree of linkage disequilibrium (data not shown). An additional dendogram using the complement set of SNPs for the 65 SNPs with significant

changes in allele frequencies was also constructed. The dendogram based on the reduced set once again resulted in the clustering of the six selected lines into a distinct cluster from the control lines ($p = 0.0011$; Figure 4-2). Further subdivision into winter and spring conditions for selected lines was no longer observed in the dendogram using the reduced dataset. The dendogram built from the complement of the 65 SNPs showing the largest differences in allele frequencies ($n = 1135$) showed that selected lines form a non-exclusive cluster with one control line inserted in it (Figure 4-3; Monte Carlo: $p = 0.0077$). Selected lines further subdivided into spring and winter clusters with the winter selected cluster containing a spring control. In both dendograms, parent lines had a higher similarity with control than selected lines. These findings show that although some SNPs show a more prominent change in allele frequencies in response to selection for early flowering, patterns of change in allele frequencies appear to be more general.

4.4.2 Parallel evolution among selected lines

Further analysis of these SNPs shows that only a few SNPs were found to have changed significantly in allele frequency in all 3 selected lines within each growth environment with 1 in winter and 2 in the spring condition associated to the same gene *GA1* (changes as large as 0.32). A further 9 SNPs identified to respond to selection in a single selected spring replicate when using population-level detection, also show significant LOD scores for the combined probabilities of all 3 selected populations changing in the same direction. In these cases, although the magnitudes of the genetic changes were clearly not replicated across all three replicated selection lines, there is also good evidence that all 3 populations show some common pattern of response. However, there are still 13 SNPs in the spring and 21 SNPs in the

winter for each there is only support for change in one selected line. For example, there is strong support for changes in the SNP LD_5903 in the winter replicate line 2 (all 3 statistical tests indicate significant changes), where the X allele went from a frequency of 0.13 to a frequency of 0.48, but no correspond increases are seen in lines 1 or 3. This indicates that the genetic basis of response to selection generally differs among replicates of the same treatment.

Given that changes in allele frequencies even after removing SNPs with the most significant changes in selected lines still show a significant difference between selected and control lines, it is possible that there is a significant degree of parallel changes in allele frequencies at the genomic scale in selected lines. We first examined the extent of parallel changes among control lines to establish the null expectation. If testing an infinite number of loci tested, then the distance separating any two daughter lines should be the sum of the distance of each line from its parent population. However, because there are only ~1200 SNPs, then a certain degree of mirroring in allele frequency change between different populations is expected, resulting in a smaller distance between them even under random genetic drift. Using the correlation coefficient matrices for all lines against each other (Table 4-2), we found that the divergence among control lines in spring conditions is slightly below than the sum of overall divergence between each population to its parent population (97.59%). We found a much stronger reduction in the extent of the divergence between selected lines compared to the sum of their divergence from the parent (Table 4-3 and Table 4-4; 70.77%). We further tested the robustness of this result by repeating the analysis in 10,000 randomly selected samples of 200 SNPs and found that the divergence among selected lines from the parent population was larger than in control lines in all cases. A similar result was obtained for the winter condition

with controls having 98.59% of the expected divergence among themselves based on the sum of their divergence from the parent line and a much reduced divergence among selected lines of 76.86% from the total expected based on the divergence of the lines from the parent population. All 10,000 random samples of 200 SNPs exhibited a lower percentage of divergence among selected lines compared to the sum of the differences against parent population compared to the case for control lines ($p > 0.0001$). These results suggest that there is a significant degree of parallel evolution among selected lines despite the fact that only a small number of SNPs show a consistent significant change in allele frequencies.

4.4.3 Limited evidence of parallel changes for winter and spring conditions

In order to explore further if changes across selected lines are independent in winter and spring conditions we test the allele frequency changes under spring and winter conditions. As is shown in Table 4-5, there are no significant differences between winter and spring conditions in terms of their divergence from parent population when examining control lines and no significant differences in the degree of divergence from parent population were observed in selected lines.

By comparing allele frequencies before and after selection, we were able to identify a number of loci that responded to selection for earlier flowering in *A. thaliana*. It is somewhat surprising that although the observed response comes from standing genetic variation, there is evidence that the response on the replicated lines exhibited large variations between different lines. While a significant degree of parallel evolution was detected among selected lines for winter and spring conditions

independently, there was little evidence of common changes in allele frequency across the two growth conditions.

These results add to the understanding of the adaptive response of allelic frequencies in response to selection of a key trait in the plant.

4.5 References

ALONSO-BLANCO, C., EL-ASSAL, S. E.-D., COUPLAND, G. & KOORNNEEF, M. 1998. Analysis of Natural Allelic Variation at Flowering Time Loci in the Landsberg erecta and Cape Verde Islands Ecotypes of Arabidopsis thaliana. *Genetics,* 149**,** 749-764.

ATWELL, S., HUANG, Y. S., VILHJALMSSON, B. J., WILLEMS, G., HORTON, M., LI, Y., MENG, D., PLATT, A., TARONE, A. M., HU, T. T., JIANG, R., MULIYATI, N. W., ZHANG, X., AMER, M. A., BAXTER, I., BRACHI, B., CHORY, J., DEAN, C., DEBIEU, M., DE MEAUX, J., ECKER, J. R., FAURE, N., KNISKERN, J. M., JONES, J. D. G., MICHAEL, T., NEMRI, A., ROUX, F., SALT, D. E., TANG, C., TODESCO, M., TRAW, M. B., WEIGEL, D., MARJORAM, P., BOREVITZ, J. O., BERGELSON, J. & NORDBORG, M. 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature,* 465**,** 627-631.

BERGELSON, J. & ROUX, F. 2010. Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nat Rev Genet,* 11**,** 867-879.

BRACHI, B., FAURE, N., HORTON, M., FLAHAUW, E., VAZQUEZ, A., NORDBORG, M., BERGELSON, J., CUGUEN, J. & ROUX, F. 2010. Linkage and Association Mapping of *Arabidopsis thaliana*Flowering Time in Nature. *PLoS Genet,* 6**,** e1000940.

BROWN, G. R., GILL, G. P., KUNTZ, R. J., LANGLEY, C. H. & NEALE, D. B. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proc Natl Acad Sci U S A,* 101**,** 15255-60.

BUCKLER, E. S., HOLLAND, J. B., BRADBURY, P. J., ACHARYA, C. B., BROWN, P. J., BROWNE, C., ERSOZ, E., FLINT-GARCIA, S., GARCIA, A., GLAUBITZ, J. C., GOODMAN, M. M., HARJES, C., GUILL, K., KROON, D. E., LARSSON, S., LEPAK, N. K., LI, H., MITCHELL, S. E., PRESSOIR, G., PEIFFER, J. A., ROSAS, M. O., ROCHEFORD, T. R., ROMAY, M. C., ROMERO, S., SALVO, S., VILLEDA, H. S., SOFIA DA SILVA, H., SUN, Q., TIAN, F., UPADYAYULA, N., WARE, D., YATES,

H., YU, J., ZHANG, Z., KRESOVICH, S. & MCMULLEN, M. D. 2009. The Genetic Architecture of Maize Flowering Time. *Science,* 325**,** 714-718.

DE MAGALHÃES, J. P. & CHURCH, G. M. 2007. Analyses of human–chimpanzee orthologous gene pairs to explore evolutionary hypotheses of aging. *Mechanisms of Ageing and Development,* 128**,** 355-364.

GALLAIS, A., MOREAU, L. & CHARCOSSET, A. 2007. Detection of marker–QTL associations by studying change in marker frequencies with selection. *Theoretical and Applied Genetics,* 114**,** 669-681.

GILKS, W. R., RICHARDSON, S. & SPIEGELHALTER, D. J. 1998. *Markov chain Monte Carlo in practice,* Boca Raton, Fla., Chapman & Hall.

GOLDRINGER, I. & BATAILLON, T. 2004. On the Distribution of Temporal Variations in Allele Frequency: Consequences for the Estimation of Effective Population Size and the Detection of Loci Undergoing Selection. *Genetics,* 168**,** 563-568.

HAERTY, W., JAGADEESHAN, S., KULATHINAL, R. J., WONG, A., RAVI RAM, K., SIROT, L. K., LEVESQUE, L., ARTIERI, C. G., WOLFNER, M. F., CIVETTA, A. & SINGH, R. S. 2007. Evolution in the Fast Lane: Rapidly Evolving Sex-Related Genes in Drosophila. *Genetics,* 177**,** 1321-1335.

JOHANSSON, A. M., PETTERSSON, M. E., SIEGEL, P. B. & CARLBORG, Ö. 2010. Genome-Wide Effects of Long-Term Divergent Selection. *PLoS Genet,* 6**,** e1001188.

KOVER, P. X., VALDAR, W., TRAKALO, J., SCARCELLI, N., EHRENREICH, I. M., PURUGGANAN, M. D., DURRANT, C. & MOTT, R. 2009a. A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in *Arabidopsis thaliana*. *PLoS Genet,* 5**,** e1000551.

Kover, P. X., J. K. Rowntree, N. Scarcelli, Y. Savriama, T. Eldridge et al., 2009b Pleiotropic effects of environment-specific adaptation in Arabidopsis thaliana. New Phytologist 183: 816-825.

LI, Y. & DE MAGALHÃES, J. 2013. Accelerated protein evolution analysis reveals genes and pathways associated with the evolution of mammalian longevity. *AGE,* 35**,** 301-314.

ROSS-IBARRA, J., MORRELL, P. L. & GAUT, B. S. 2007. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc Natl Acad Sci U S A,* 104 Suppl 1**,** 8641-8.

SCARCELLI, N., CHEVERUD, J. M., SCHAAL, B. A. & KOVER, P. X. 2007. Antagonistic pleiotropic effects reduce the potential adaptive value of the FRIGIDA locus. *Proc Natl Acad Sci U S A,* 104**,** 16986-91.

SCARCELLI, N. & KOVER, P. X. 2009. Standing genetic variation in FRIGIDA mediates experimental evolution of flowering time in Arabidopsis. *Molecular Ecology,* 18**,** 2039-2049.

TURNER, T. L., STEWART, A. D., FIELDS, A. T., RICE, W. R. & TARONE, A. M. 2011. Population-Based Resequencing of Experimentally Evolved Populations Reveals the Genetic Basis of Body Size Variation in *Drosophila melanogaster*. *PLoS Genet,* 7**,** e1001336.

UNGERER, M. C., HALLDORSDOTTIR, S. S., PURUGGANAN, M. D. & MACKAY, T. F. C. 2003. Genotype-Environment Interactions at Quantitative Trait Loci Affecting Inflorescence Development in Arabidopsis thaliana. *Genetics,* 165**,** 353-365.

WEINIG, C., DORN, L. A., KANE, N. C., GERMAN, Z. M., HALLDORSDOTTIR, S. S., UNGERER, M. C., TOYONAGA, Y., MACKAY, T. F. C., PURUGGANAN, M. D. & SCHMITT, J. 2003. Heterogeneous Selection at Specific Loci in Natural Environments in Arabidopsis thaliana. *Genetics,* 165**,** 321-329.

WILCZEK, A. M., ROE, J. L., KNAPP, M. C., COOPER, M. D., LOPEZ-GALLEGO, C., MARTIN, L. J., MUIR, C. D., SIM, S., WALKER, A., ANDERSON, J., EGAN, J. F., MOYERS, B. T., PETIPAS, R., GIAKOUNTIS, A., CHARBIT, E., COUPLAND, G., WELCH, S. M. & SCHMITT, J. 2009. Effects of Genetic Perturbation on Seasonal Life History Plasticity. *Science,* 323**,** 930-934.

ZHANG, J. H. & LECHOWICZ, M. J. 1994. Correlation between Time of Flowering and Phenotypic Plasticity in Arabidopsis-Thaliana (Brassicaceae). *American Journal of Botany,* 81**,** 1336-1342.

4.6 Tables and figures

Table 4-21.Number of SNPS showing significant changes in allele frequencies in control and selected lines, grown in winter and spring treatments, using four different methods.

	Gallais X2		Fc		Markov Chain		Cumulative	
	Control	Select.	Control	Select.	Control	Select.	Control	Select.
Spring								
Line 1	2	22	26 (9)	61	6	15		
Line 2	0	4	0	6	0	2	2	22
Line 3	0	0	0	2	0	0		
Winter								
Line 1	1	24	16	46	2	19		
Line 2	0	8	0	11	0	3	4	20
Line 3	0	1	0	1	0	1		

Table 4-22.Correlation coefficients among all parent, control and selected lines.

	SG5C1	SG5C2	SG5C3	SG5S1	SG5S2	SG5S3	WG5C1	WG5C2	WG5C3	WG5S1	WG5S2
SG5C1											
SG5C2	0.912										
SG5C3	0.931	0.923									
SG5S1	0.911	0.908	0.920								
SG5S2	0.912	0.891	0.913	0.927							
SG5S3	0.911	0.876	0.909	0.917	0.933						
WG5C1	0.940	0.925	0.941	0.926	0.919	0.919					
WG5C2	0.935	0.917	0.933	0.914	0.905	0.904	0.943				
WG5C3	0.927	0.915	0.931	0.907	0.906	0.906	0.931	0.924			
WG5S1	0.910	0.895	0.927	0.895	0.884	0.888	0.924	0.920	0.902		
WG5S2	0.905	0.884	0.916	0.915	0.906	0.907	0.921	0.920	0.909		
WG5S3	0.911	0.908	0.924	0.911	0.910	0.893	0.927	0.924	0.914	0.922	0.912

Table 4-23.Distance of each line from parent populations within spring and winter.

Line A	Line B	Percentage of line B larger than line A (%)
Spring		
C[1] 1	C 2	99.72
C 1	C 3	3.39
C 2	C 3	0
S[2] 1	S 2	67.08
S 1	S 3	61.66
S 2	S 3	44.62
Winter		
C 1	C 2	97.04
C 1	C 3	99.7
C 2	C 3	80.72
S 1	S 2	30.95
S 1	S 3	2.48
S 2	S 3	10.78

[1] Control line

[2] Selected line

Table 4-24.Divergence between control lines and selected lines within spring and winter.

Sprint line A	Spring line B	Percentage of line B larger than line A (%)	Winter line A	Winter line B	Percentage of line B larger than line A (%)
C[1] 1	S[2] 1	99.03	C 1	S 1	100
C 1	S 2	99.67	C 1	S 2	100
C 1	S 3	99.5	C 1	S 3	100
C 2	S 1	33.68	C 2	S 1	100
C 2	S 2	48.41	C 2	S 2	100
C 2	S 3	43.87	C 2	S 3	99.34
C 3	S 1	100	C 3	S 1	99.96
C 3	S 2	100	C 3	S 2	99.69
C 3	S 3	100	C 3	S 3	93.41

[1] Control line

[2] Selected line

Table 4-25. Difference resulted by spring and winter conditions.

Sprint line	Winter line	Percentage of winter larger than spring (%)	Spring line	Winter line	Percentage of winter larger than spring (%)
C[1] 1	C 1	99.03	S[2] 1	S 1	100
C 1	C 2	99.67	S 1	S 2	100
C 1	C 3	99.5	S 1	S 3	100
C 2	C 1	33.68	S 2	S 1	100
C 2	C 2	48.41	S 2	S 2	100
C 2	C 3	43.87	S 2	S 3	99.34
C 3	C 1	100	S 3	S 1	99.96
C 3	C 2	100	S 3	S 2	99.69
C 3	C 3	100	S 3	S 3	93.41

[1] Control line

[2] Selected line

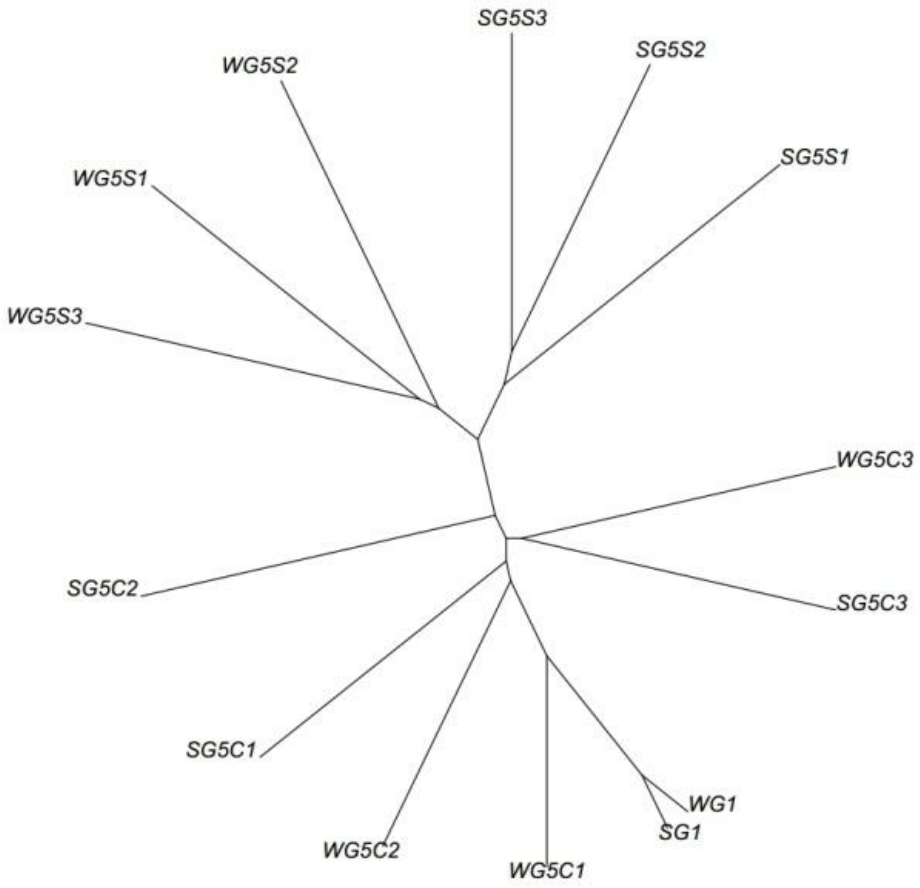


Figure 4-31.Dendrogram of spring/winter lines based on allel frequencies before and after selection in six generations

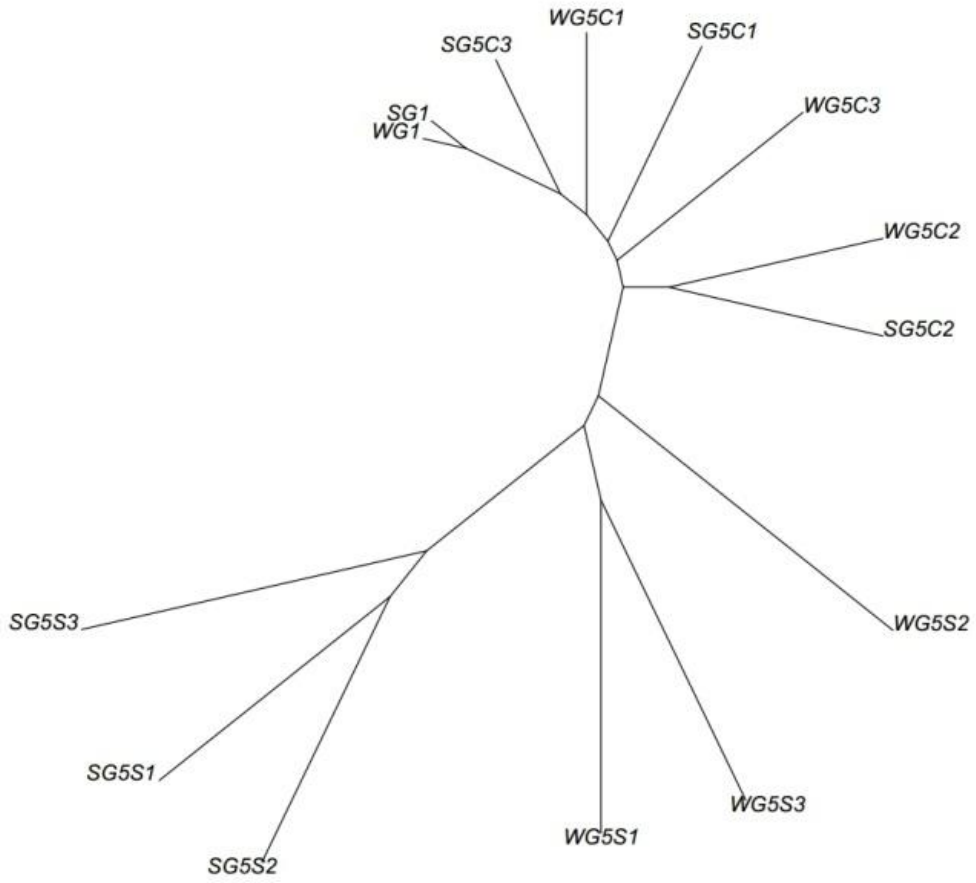


Figure 4-32.Dendrogram of spring/winter lines based on allele frequencies before and after selection in six generations with 65 (spring/winter) SNPs.

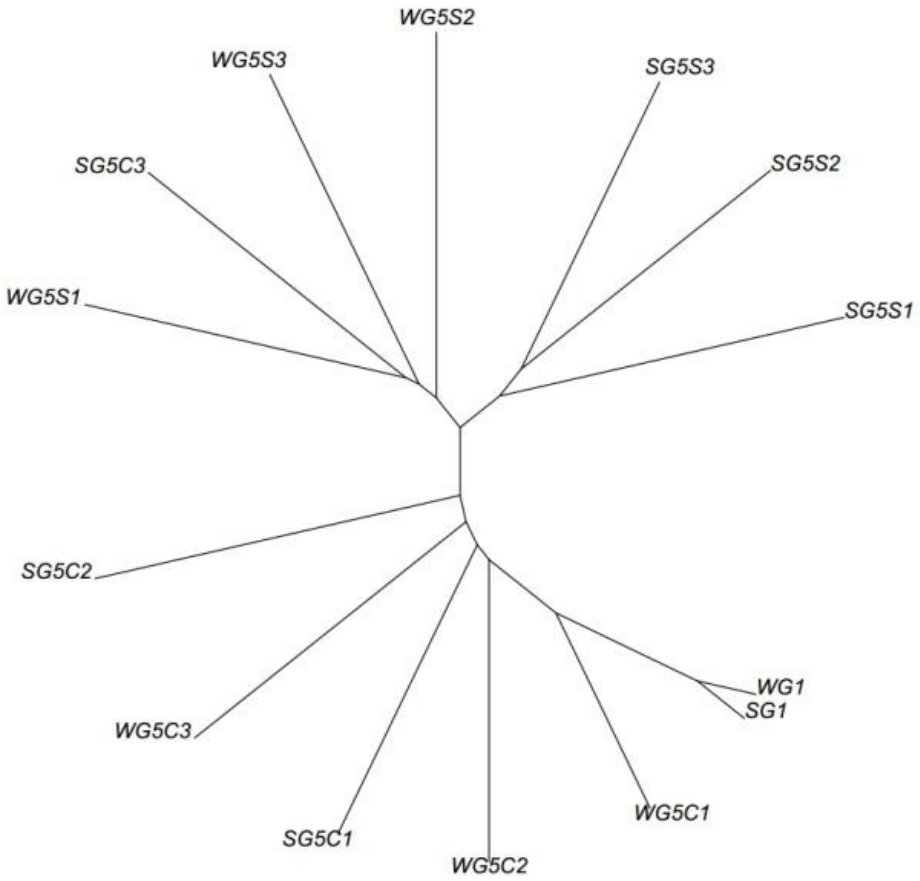


Figure 4-33.Dendrogram of spring/winter lines based on allele frequencies before and after selection in six generations without 65 (spring/winter) SNPs.

4.7 Supplementary tables and figures

Supplementary table S4-26.Correlation between control and selected lines.

	Correlation coefficient	p values
Winter	0.9667794	< 2.2e-16
Spring	0.9542208	< 2.2e-16

Supplementary table S4-27. Distance matrix for all SNPs loci.

	SP[1]	SC[2]1	SC2	SC3	SS[3]1	SS2	SS3	WP[4]	WC[5]1	WC2	WC3	WS[6]1	WS2
SC1	0.276												
SC2	0.321	0.410											
SC3	0.248	0.366	0.386										
SS1	0.319	0.412	0.418	0.393									
SS2	0.326	0.409	0.455	0.407	0.376								
SS3	0.328	0.412	0.483	0.417	0.398	0.360							
WG1	0.066	0.279	0.317	0.250	0.320	0.327	0.328						
WC1	0.238	0.340	0.379	0.339	0.377	0.394	0.395	0.233					
WC2	0.267	0.356	0.398	0.359	0.406	0.425	0.427	0.266	0.333				
WC3	0.290	0.374	0.403	0.364	0.421	0.424	0.423	0.281	0.365	0.381			
WS1	0.346	0.415	0.446	0.375	0.446	0.467	0.460	0.344	0.383	0.392	0.433		
WS2	0.338	0.426	0.468	0.401	0.403	0.422	0.421	0.341	0.389	0.392	0.416	0.410	
WS3	0.309	0.413	0.419	0.382	0.412	0.414	0.449	0.310	0.375	0.382	0.407	0.387	0.411

[1] SP: spring parent
[2] SC: spring control
[3] SS: spring selected
[4] WP: winter parent
[5] WC: winter control
[6] WS: winter selected

Supplementary table S4-28. Distance matrix for 65 SNPs loci.

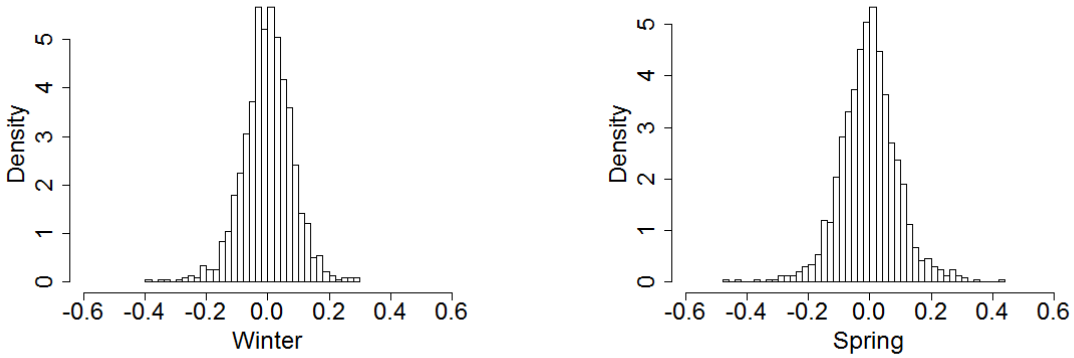
	SP[1]	SC[2]1	SC2	SC3	SS[3]1	SS2	SS3	WP[4]	WC[5]1	WC2	WC3	WS[6]1	WS2
SC1	0.279												
SC2	0.305	0.404											
SC3	0.234	0.344	0.311										
SS1	0.570	0.613	0.642	0.608									
SS2	0.567	0.563	0.648	0.628	0.485								
SS3	0.600	0.607	0.721	0.666	0.582	0.542							
WG1	0.073	0.285	0.311	0.245	0.573	0.561	0.592						
WC1	0.250	0.356	0.358	0.299	0.560	0.574	0.589	0.243					
WC2	0.319	0.417	0.361	0.316	0.603	0.645	0.698	0.311	0.366				
WC3	0.312	0.377	0.436	0.382	0.667	0.641	0.657	0.309	0.391	0.401			
WS1	0.648	0.674	0.613	0.577	0.729	0.790	0.793	0.649	0.608	0.600	0.643		
WS2	0.599	0.619	0.613	0.584	0.624	0.644	0.728	0.608	0.592	0.573	0.583	0.629	
WS3	0.546	0.598	0.523	0.533	0.630	0.649	0.740	0.555	0.544	0.552	0.610	0.548	0.566

[1] SP: spring parent
[2] SC: spring control
[3] SS: spring selected
[4] WP: winter parent
[5] WC: winter control
[6] WS: winter selected

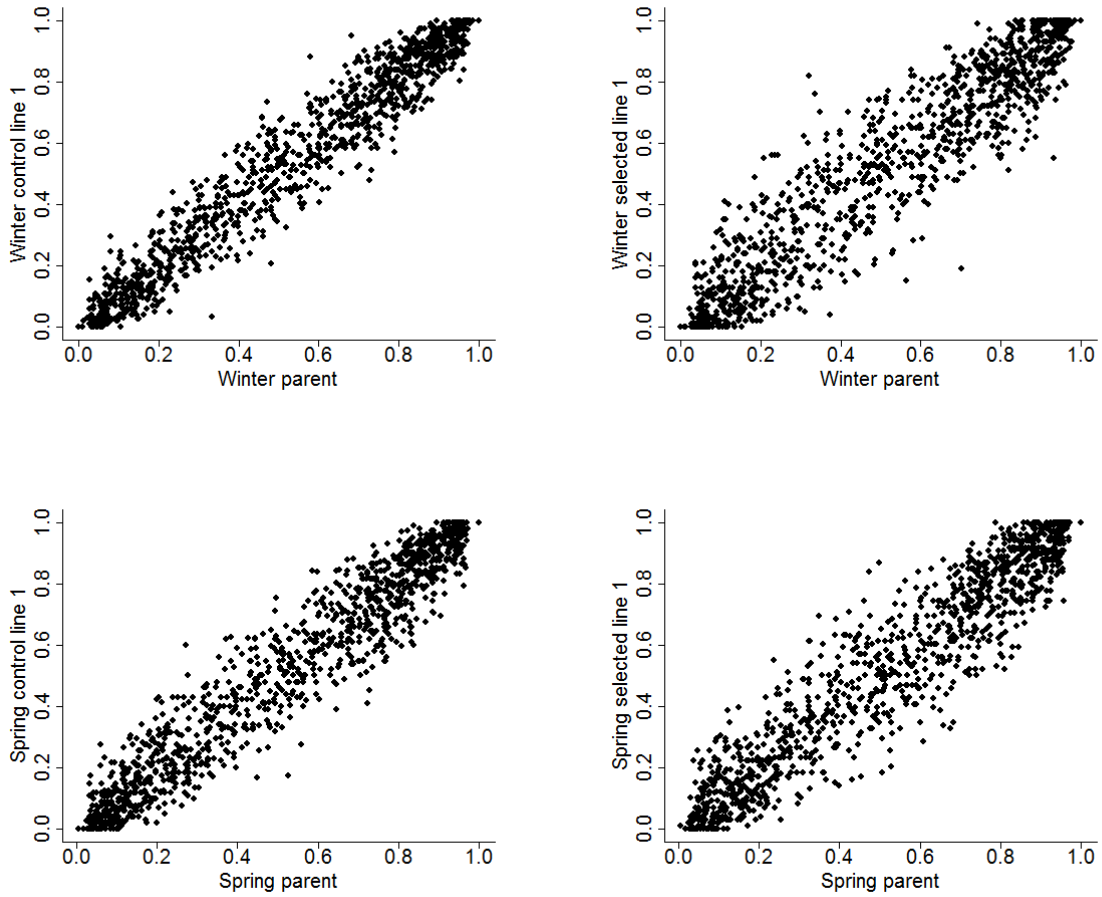
Supplementary table S4-29. Distance matrix for loci without 65 SNPs.

	SP[1]	SC[2]1	SC2	SC3	SS[3]1	SS2	SS3	WP[4]	WC[5]1	WC2	WC3	WS[6]1	WS2
SC1	0.276												
SC2	0.322	0.410											
SC3	0.249	0.366	0.389										
SS1	0.298	0.397	0.400	0.378									
SS2	0.307	0.399	0.439	0.391	0.369								
SS3	0.305	0.398	0.463	0.398	0.387	0.349							
WG1	0.066	0.278	0.317	0.250	0.299	0.309	0.307						
WC1	0.238	0.339	0.379	0.341	0.365	0.382	0.382	0.232					
WC2	0.264	0.352	0.399	0.361	0.393	0.410	0.408	0.263	0.331				
WC3	0.289	0.374	0.401	0.363	0.401	0.408	0.405	0.279	0.363	0.380			
WS1	0.320	0.393	0.432	0.361	0.427	0.442	0.434	0.317	0.367	0.378	0.416		
WS2	0.316	0.411	0.457	0.389	0.390	0.407	0.398	0.319	0.375	0.380	0.404	0.397	
WS3	0.288	0.399	0.412	0.371	0.397	0.396	0.426	0.289	0.364	0.370	0.391	0.376	0.401

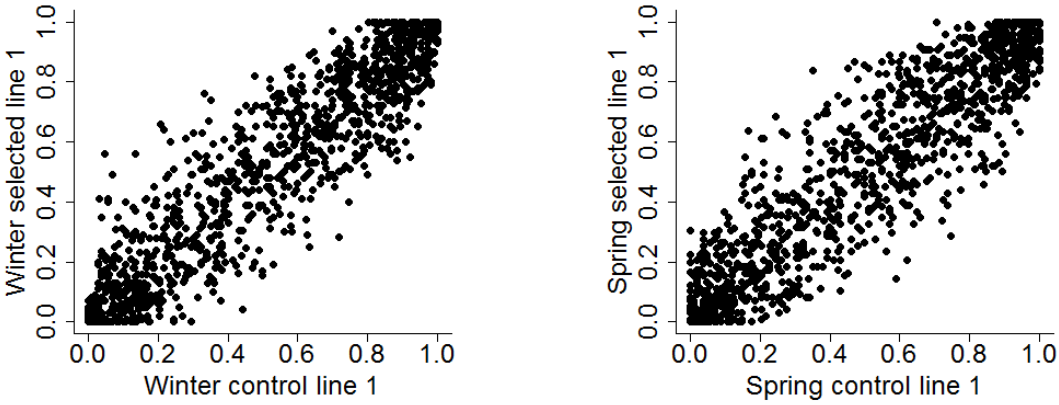
[1] SP: spring parent
[2] SC: spring control
[3] SS: spring selected
[4] WP: winter parent
[5] WC: winter control
[6] WS: winter selected



Supplementary figure S4-34. Histogram of differences between averaged control lines and averaged selected lines.



Supplementary figure S4-35. Dot plots of parents against control/selected lines.



Supplementary figure S4-36. Dot plots of winter/spring control lines against winter/spring selected lines.

Chapter 5 Discussion

Genome organisation and the rearrangements it undergoes across time, as well as the patterns of change in response to selective pressures, are of great importance to our understanding of the factors driving the evolution of genomes and adaptation. In this thesis I have examined different aspects of genome organization.Changes in DNA sequences are examined in order to identify signatures of adaptive evolution on a genome wide scale. In each chapter, I characterized patterns of genome organisation according to patterns of sex biased gene expression in primate species and testis overexpression in the Drosophila lineage, as well as allele frequency changes in the model plant *Arabidopsis thaliana* in response to selective pressures.

5.1 Sex-biased genes in primates

In chapter 2, I examined patterns of sex biased gene expression on a genomic scale using transcriptome data available for six primate species in six different tissues. Sex biased gene expression has been reported in several species and is likely to underlie the physiological and behavioural differences between males and females. My results confirm several previously reported patterns for sex biased genes in previous studies examining single species or multiple species from other taxa. A comparative analysis in the Drosophila lineage demonstrated that male-biased expression is greater than female-biased expression in 5 out of 7 species (ZHANG *et al.* 2007). Similar results were obtained in our study of six primate species. We

observed a higher number of genes with male-biased expression across primate species except macaque.

The breadth of expression can be different for individual genes with the range from broad (genes expressed in multiple tissues) to narrow (genes expressed in specific tissues). In Drosophila, sex-biased genes have a tendency to be narrowly expressed in a limited number of tissues (MEISEL *et al.* 2012). In accordance with the above, our analysis shows that, in primate, gene with higher sex-biased expression is also associated with higher tissue specificity. A recent study has shown that in both *Drosophila melanogaster* and *Drosophila pseudoobscura*, genes with female-biased expression tend to have smaller tau values than genes with male-biased expression (ASSIS *et al.* 2012), indicating that female-biased genes show greater pleiotropy. In contrast to what has been found in Drosophila, in primates, we observed weak evidence that female-biased genes show greater tau values in gorilla, bonobo and human.

Not all of our results are consistent with patterns previously shown in other studies. Genes with sex-biased expression, especially male-biased genes, tend to rapidly evolve in both DNA sequence and expression level (RANZ *et al.* 2003; ZHANG *et al.* 2004). Nevertheless, male-biased gene expression contributes greatly to overall expression divergence (ZHANG *et al.* 2007). In Drosophila, sex-biased genes, particularly those expressed in reproductive tissues, show elevated levels of amino acid divergence between species (PROSCHEL *et al.* 2006; ZHANG *et al.* 2004). Male-biased genes often demonstrate evidence of higher divergence rates than female-biased genes between species (ELLEGREN and PARSCH 2007; ZHANG *et al.* 2007). We used synonymous and non-synonymous rates in protein coding genes to assess DNA

sequence divergence in primates, and, in contrast, the results did not show significant associations between sex-biased genes and dN/dS across six primate species. However, we did not observe significant differences of dN/dS between female- and male-biased genes (Supplementary table S2-10; Supplementary figureS2-7 – S2-10). We are not sure about the possible causes of this difference. It is possible for example that the divergence times from mouse to the primate species examinedis too high to allow accurate estimates of dN/dS.

Interestingly, our results provide the first evidence suggesting significant clustering of sex biased gene expression within chromosomes. Moreover, a gene neighbourhood affect, by which two female-/male- biased genes tend to be together, is observed. It is of course possible that this observation might be caused by the cluster of testis genes along chromosome. As a step towards addressing this question, we examined the correlation between testis gene and sex-biased gene cluster. The regression analysis indicated that testis genes did not contribute to the clustering of sex-biased genes. An alternative explanation for the cluster might be small-scale rearrangement or sexual antagonism, and this remains to be addressed in future studies.This clustering of sex biased genes could result from spill over bias to surrounding genes for a handful of genes with functional sexually dimorphic expression either through chromatin structure or co-regulation and raises the possibility that the high prevalence of sex bias expression and the poor conservation of this biases across species at a genomic scale may reflect that sex biased gene expression patterns is to some extent non-functional. Previous studies had shown an unequal distribution of sex biased genes among autosomes and sex chromosomes. Genes with female-/male-biased expression are distributed non-randomly between autosomes and sex chromosomes (ELLEGREN and PARSCH 2007; MANK and

ELLEGREN 2009; MUELLER *et al.* 2008; NAURIN *et al.* 2011; PARISI *et al.* 2003; RANZ *et al.* 2003; REINKE *et al.* 2000) but clustering of sex biased gene expressed genes within chromosomes has not been explored.

5.2 Gene order evolution of Drosophila testis over-expressed genes

Genomic rearrangement may compose an important part in the evolution of local adaptation and genomic divergence (YEAMAN 2013).In eukaryotic genomes, several forces may contribute to the formation of gene cluster. SEOIGHE *et al.*(2000) compared two genomes: *Saccharomyces cerevisiae* and *Candida albicans*. They found only 9% of adjacent gene pairs are conserved between these two species. Seoighe and colleagues (2000) revealed that small segment DNA inversion has been a major cause of rearrangement. Moreover, even in a conserved gene pair, the transcriptional orientations tend to be different between species. They further proposed that the frequency of linkage breakage of two genes by local rearrangements is as much as by long-distance transpositions or by chromosomal translocations.

It has been observed in yeast that essential gene clusters locate in low recombination regions along the chromosomes (PAL and HURST 2003), which indicates a co-evolution of recombination rate and gene order. Using high-resolution chromosome rearrangement data, WEBER and HURST(2011) examined the gene order conservation in *Drosophila melanogaster*, and the result they obtained also shows that gene order rearrangement is associated with recombination rate.

In Drosophila, male expressed genes tend to under-represented on the X chromosome, and accumulate on autosomes (DORUS *et al.* 2006; PARISI *et al.* 2004;

PARISI *et al.* 2003; RANZ *et al.* 2003). In particular, testis expressed genes have been found to be clustered in Drosophila (BOUTANAEV *et al.* 2002) and testis expressed genes have been observed to be frequently retroposed from the X chromosome to autosomes (BAI *et al.* 2007; BETRAN *et al.* 2002; DAI *et al.* 2006).

Previous studies demonstrate that the paucity of X-linked testis-expressed *D. melanogaster* genes may be limited to genes expressed during later stages of spermatogenesis such as meiosis(GAN *et al.* 2010; VIBRANOVSKI *et al.* 2009). In mammals, more cell divisions are required in spermatogenesis than in oogenesis. Therefore, higher mutation rate is expected in the male germ line than in the female germ line (VICOSO and CHARLESWORTH 2006). Baines *et al*(2008) proposed that the X-linked genes differ from the autosomal genes in their function or in the timing of their expression (e.g., early vs. late spermatogenesis).

We conducted a comparative analysis in Drosophila to examine whether the testis expressed gene clusters are more likely to be conserved. We assessed the conservation of testis gene pairs in Drosophila genus. Our results show that testis overexpressed gene pairs observed in *Drosophila melanogaster*are more likely to be unlinked in other species. Meanwhile, the transcription direction of testis gene pairs is also not conserved. By tracing back the formation of these testis clusters in evolution, we notice that the higher proportion of linkage breaks may be a result of chromosomal rearrangements. We do, however, find that genes with testis expression in *Drosophila melanogaster* existed early through evolution. We also verified that the linkage of two testis gene pairs was acquired late. This could be interpreted by a non-testis gene firstly rearranged next to a testis and then acquired its testis function, or the other way round. When we tracked the testis gene pairs in

Drosophila, we discovered that linkage break is higher for novel testis gene pairs. Then we consider whether a novel testis gene acquired its testis expression by relocating next to a conserved testis gene, we find a tendency of novel testis gene moving to a conserved testis gene. Our results on the conservation of testis clusters are consistent with previous results suggesting a poor conservation of gene clusters according to other parameters of gene expression profiles (WEBER and HURST 2011) and overall suggest that the formation of testis overexpressed gene clusters is probably the result of non-specific increases in testis expression of neighbouring genes.

5.3 Allele frequency changes in *Arabidopsis thaliana*

In the previous chapters, I assessed the evolution patterns of sex-biased gene divergence between primates, as well as the testes genes between 11 Drosophila species. Instead of investigating the evolutionary processes in real time, those studies only reflect past evolution. And we cannot confine the conditions (environmental or demographic) during evolutionary changes. Obviously, evolutionary questions cannot be addressed by a single model system or species.

Fuelled by the sufficient development of molecular methods, we can now use experimental evolution method to study the molecular basis of phenotypic changes occurring during evolution. Nevertheless, *Arabidopsis thaliana* has a diverse array of genetic resources, and it provides an excellent short-term model system to address certain fundamental evolutionary questions.

Changes to the phenotype reflect underlying changes to the genotype but despite the increasing availability of fully sequenced genomes and

transcriptomesdata for an increasing number of species, establishing links between specific changes at the sequence level and or changes in gene expression profiles to complex phenotypes such as adaptation to climate changes remains difficult. Using genome wide analysis, I investigate molecular changes that mediate short-term response to selection for earlier flowering time in *Arabidopsis thaliana* under two environmental conditions.We found that control and selected lines formed separate clusters, and the cluster of parent population suggests that control lines diverged less in their allele frequencies compared to the selected lines. Meanwhile, although some SNPs show a more prominent change in allele frequencies in response to selection for early flowering, patterns of change in allele frequencies appear to be more general. Selected lines exhibit significantly larger differences in allele frequencies with respect to the parent population compared to control lines. We further show that,despite the fact that only a small number of SNPs show a consistent significant change in allele frequencies, allele distributions are consistent with significant parallel evolution among replica lines.

A previous study in domesticated chicken has investigated genome-wide long term responses to selection for body weight in 50 generations (JOHANSSON *et al.* 2010). The selection effect was dramatic on the phenotype. They also discovered >100 regions with different genetic variants between two lines. The significant differences identified in many regions scattered over the entire genome. However, the SNPs are not randomly distributed.

5.4 General conclusion

In this thesis, firstly, I characterised sex-biased gene expression in five tissues for of six primate species. We showed that higher levels of sex-biased expression are associated with lower expression and higher tissue specificity. In addition, we found that sex-biased genes tend to cluster in specific regions within chromosomes which suggests a strong influence of genomic neighbourhoods and chromatin structure in defining levels of sex-biased gene expression.

Furthermore, we assessed the conservation of testis gene clusters. The result shows that *Drosophila melanogaster* testis clusters are less likely to be found linked in other Drosophila species. This observation can be explained by both a higher rate of linkage breakage and the later linkage acquisition for genes not ancestrally linked. New neighbours for ancestrally linked *Drosophila melanogaster* testis genes tend to be other testis expressed genes but this is not statistically significant.

As testis genes are rapidly evolved, it is hard to detect the orthologous across species even in closely related species such as in Drosophila genus. Therefore, we, alternatively, used male-biased data together with our testis expression data to trace the ancestral gene with testis expression. Nevertheless, with the lack of expression data of different spermatogenesis stages, we failed to assess details of the relation between testis gene pair linkage and spermatogenesis.

Finally, using an experimental evolution method, I assessed the selection acting on allele frequencies and genomic organisation through the analysis of various genome-wide data from *Arabidopsis thaliana* to the genus of Drosophila to six primate species. We found that selection for early flowering resulted in higher allele

frequency divergence from parent populations. Meanwhile, selected lines showed parallel changes in allele frequencies, but selection for early flowering was independent in both winter and spring conditions. Experimental evolution study of allele frequency changes in *Arabidopsis thaliana* provides an alternative approach to identifying the genetic basis of natural variation in complex traits.

5.5 Future studies

Various questions could be addressed in the future following the work presented in the thesis.

With the data increasing, in the future we could compare break rates for genes expressed at different stages of spermatogenesis. It is also worth to explore whether these findings are particular to testis gene clusters or also apply to other tissue specific genes.

At the genomic level, what underlies behavioural differences between females and males in primate remains less understood. We could investigate transcriptome differences between males and females by using data such as body mass, and identify genes whose expression correlates with the degree of monogamy/polygamy.

As for continuing the assessment of the allele frequency changes in Arabidopsis, it would be interesting to make a systematic assessment of the clustering of SNPs showing the highest rates of change in the selected lines. A functional characterisation of the neighbouring genes to the SNPs of interest would also be interesting as it could reveal further consistency in changes in allele

frequency in genes associated with distinct functions even if no consistency is observed across individual SNPs in replica lines.

5.6 References

ASSIS, R., Q. ZHOU and D. BACHTROG, 2012 Sex-biased transcriptome evolution in Drosophila. Genome Biol Evol **4:** 1189-1200.

BAI, Y., C. CASOLA, C. FESCHOTTE and E. BETRAN, 2007 Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. Genome Biology **8:** R11.

BAINES, J. F., S. A. SAWYER, D. L. HARTL and J. PARSCH, 2008 Effects of X-linkage and sex-biased gene expression on the rate of adaptive protein evolution in Drosophila. Mol Biol Evol **25:** 1639-1650.

BETRAN, E., K. THORNTON and M. LONG, 2002 Retroposed new genes out of the X in Drosophila. Genome Res **12:** 1854-1859.

BOUTANAEV, A. M., A. I. KALMYKOVA, Y. Y. SHEVELYOV and D. I. NURMINSKY, 2002 Large clusters of co-expressed genes in the Drosophila genome. Nature **420:** 666-669.

DAI, H., T. F. YOSHIMATSU and M. LONG, 2006 Retrogene movement within- and between-chromosomes in the evolution of Drosophila genomes. Gene **385:** 96-102.

DORUS, S., S. A. BUSBY, U. GERIKE, J. SHABANOWITZ, D. F. HUNT *et al.*, 2006 Genomic and functional evolution of the Drosophila melanogaster sperm proteome. Nat Genet **38:** 1440-1445.

ELLEGREN, H., and J. PARSCH, 2007 The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet **8:** 689-698.

GAN, Q., I. CHEPELEV, G. WEI, L. TARAYRAH, K. CUI *et al.*, 2010 Dynamic regulation of alternative splicing and chromatin structure in Drosophila gonads revealed by RNA-seq. Cell Res **20:** 763-783.

JOHANSSON, A. M., M. E. PETTERSSON, P. B. SIEGEL and Ö. CARLBORG, 2010 Genome-Wide Effects of Long-Term Divergent Selection. PLoS Genet **6:** e1001188.

MANK, J. E., and H. ELLEGREN, 2009 Sex-linkage of sexually antagonistic genes is predicted by female, but not male, effects in birds. Evolution **63:** 1464-1472.

MEISEL, R. P., J. H. MALONE and A. G. CLARK, 2012 Disentangling the relationship between sex-biased gene expression and X-linkage. Genome Res **22:** 1255-1265.

MUELLER, J. L., S. K. MAHADEVAIAH, P. J. PARK, P. E. WARBURTON, D. C. PAGE *et al.*, 2008 The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. Nat Genet **40:** 794-799.

NAURIN, S., B. HANSSON, D. HASSELQUIST, Y. H. KIM and S. BENSCH, 2011 The sex-biased brain: sexual dimorphism in gene expression in two species of songbirds. BMC Genomics **12:** 37.

PAL, C., and L. D. HURST, 2003 Evidence for co-evolution of gene order and recombination rate. Nat Genet **33:** 392-395.

PARISI, M., R. NUTTALL, P. EDWARDS, J. MINOR, D. NAIMAN *et al.*, 2004 A survey of ovary-, testis-, and soma-biased gene expression in Drosophila melanogaster adults. Genome Biology **5:** R40.

PARISI, M., R. NUTTALL, D. NAIMAN, G. BOUFFARD, J. MALLEY *et al.*, 2003 Paucity of genes on the Drosophila X chromosome showing male-biased expression. Science **299:** 697-700.

PROSCHEL, M., Z. ZHANG and J. PARSCH, 2006 Widespread adaptive evolution of Drosophila genes with sex-biased expression. Genetics **174:** 893-900.

RANZ, J. M., C. I. CASTILLO-DAVIS, C. D. MEIKLEJOHN and D. L. HARTL, 2003 Sex-dependent gene expression and evolution of the Drosophila transcriptome. Science **300:** 1742-1745.

REINKE, V., H. E. SMITH, J. NANCE, J. WANG, C. VAN DOREN *et al.*, 2000 A Global Profile of Germline Gene Expression in C. elegans. Molecular Cell **6:** 605-616.

SEOIGHE, C., N. FEDERSPIEL, T. JONES, N. HANSEN, V. BIVOLAROVIC *et al.*, 2000 Prevalence of small inversions in yeast gene order evolution. Proc Natl Acad Sci U S A **97:** 14433-14437.

VIBRANOVSKI, M. D., H. F. LOPES, T. L. KARR and M. LONG, 2009 Stage-specific expression profiling of Drosophila spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. PLoS Genet **5:** e1000731.

VICOSO, B., and B. CHARLESWORTH, 2006 Evolution on the X chromosome: unusual patterns and processes. Nat Rev Genet **7:** 645-653.

WEBER, C. C., and L. D. HURST, 2011 Support for multiple classes of local expression clusters in Drosophila melanogaster, but no evidence for gene order conservation. Genome Biol **12:** R23.

YEAMAN, S., 2013 Genomic rearrangements and the evolution of clusters of locally adaptive loci. Proc Natl Acad Sci U S A **110:** E1743-1751.

ZHANG, Y., D. STURGILL, M. PARISI, S. KUMAR and B. OLIVER, 2007 Constraint and turnover in sex-biased gene expression in the genus Drosophila. Nature **450:** 233-237.

ZHANG, Z., T. M. HAMBUCH and J. PARSCH, 2004 Molecular evolution of sex-biased genes in Drosophila. Mol Biol Evol **21:** 2130-2139.